Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective *

Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn and Linda Zhao

November 9, 2002

Corresponding author:
Lawrence D. Brown
Department of Statistics, The Wharton School, University of Pennsylvania,
Philadelphia, PA 19104-6340
Email: lbrown@wharton.upenn.edu, Phone: (215)898-4753, Fax: (215)898-1280

*Lawrence Brown is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, (email: lbrown@wharton.upenn.edu). Noah Gans is Assistant Professor, Department of Operations and Information Management, The Wharton School, University of Pennsylvania, (email: gans@wharton.upenn.edu). Avishai Mandelbaum is Professor, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, (email: avim@techunix.technion.ac.il). Anat Sakov is post-doctoral fellow, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, (email: sakov@ie.technion.ac.il). Haipeng Shen is Ph.D. Candidate, Department of Statistics, The Wharton School, University of Pennsylvania, (email: haipengs@wharton.upenn.edu). Sergey Zeltyn is Ph.D. Candidate, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, (email: zeltyn@ie.technion.ac.il). Linda Zhao is Associate Professor, Department of Statistics, The Wharton School, University of Pennsylvania, (email: lzhao@wharton.upenn.edu). This work was supported by the ISF (Israeli Science Foundation) Grants 388/99 and 126/02, the Wharton Financial Institutions Center, the NSF and the Technion funds for the promotion of research and sponsored research.

Abstract

A call center is a service network in which agents provide telephone-based services. Customers that seek these services are delayed in tele-queues.

This paper summarizes an analysis of a unique record of call center operations. The data comprise a complete operational history of a small banking call center, call by call, over a full year. Taking the perspective of queueing theory, we decompose the service process into three fundamental components: arrivals, customer abandonment behavior and service durations. Each component involves different basic mathematical structures and requires a different style of statistical analysis. Some of the key empirical results are sketched, along with descriptions of the varied techniques required.

Several statistical techniques are developed for analysis of the basic components. One of these is a test that a point process is a Poisson process. Another involves estimation of the mean function in a nonparametric regression with lognormal errors. A new graphical technique is introduced for nonparametric hazard rate estimation with censored data. Models are developed and implemented for forecasting of Poisson arrival rates.

We then survey how the characteristics deduced from the statistical analyses form the building blocks for theoretically interesting and practically useful mathematical models for call center operations.

Key Words: call centers, queueing theory, lognormal distribution, inhomogeneous Poisson process, censored data, human patience, prediction of Poisson rates, Khintchine-Pollaczek formula, service times, arrival rate, abandonment rate, multiserver queues.

Contents

1 INTRODUCTION **1**

1.1 Highlights of results 2

1.2 Structure of the paper 4

2 QUEUEING MODELS OF CALL CENTERS **4**

3 THE CALL CENTER OF BANK ANONYMOUS **5**

4 THE ARRIVAL PROCESS **7**

4.1 Arrivals are inhomogeneous Poisson 9

4.2 The Poisson arrival-rates are not easily "predictable" 13

5 SERVICE TIME **15**

5.1 Very short service times 16

5.2 On service times and queueing theory 17

5.3 Service times are lognormal 18

5.4 Regression of log service times on time-of-day 19

5.4.1 Estimation of $\mu(\cdot)$ 19

5.4.2 Estimation of $\sigma^2(\cdot)$ 20

5.4.3 Estimation of $\nu(\cdot)$ 20

5.4.4 Application and model diagnostics 21

6 QUEUEING TIME: WAITING TIME FOR SERVICE OR ABANDONING **23**

6.1 Waiting times are exponentially distributed 23

6.2 Survival curves for virtual waiting time and patience 24

6.3 Hazard rates 27

6.4 Patience Index 29

6.5 Dependence, or the violation of classical assumptions of survival analysis 31

7 PREDICTION OF THE LOAD 32

7.1 Definition of load 32

7.2 Independence of $\Lambda(t)$ and $\nu(t)$ 33

7.3 Coefficient of Variation for the prediction of $L(t)$ 34

7.4 Prediction of $\Lambda(t)$ 35

7.5 Diagnostics for the model for $\Lambda(t)$ 38

7.6 Prediction of $\nu(t)$ 38

7.7 Confidence intervals for $L(t)$ 40

8 SOME APPLICATIONS OF QUEUEING SCIENCE 41

8.1 Validating Classical Queueing Theory 42

8.2 Fitting the M/M/N+M model (Erlang-A) 46

8.2.1 The Erlang-A Model 47

8.2.2 Approximations 49

8.2.3 Use and Limits of the Erlang-A Model 50

9 CONCLUSION 51

1 INTRODUCTION

Telephone call centers allow groups of agents to serve customers remotely, via the telephone. They have become a primary contact point between customers and their service providers and, as such, play an increasingly significant role in more developed economies. For example, it is estimated that call centers handle more than 70% of all business interactions and that they employ more than 3.5 million people, or 2.5% of the workforce, in the U.S. (Uchitelle, 2002; Call Center Statistics, 2002).

While call centers are technology-intensive operations, often 70% or more of their operating costs are devoted to human resources, and to minimize costs their managers carefully track and seek to maximize agent utilization. Well-run call centers adhere to a sharply-defined balance between agent efficiency and service quality, and to do so, they use queueing-theoretic models. Inputs to these mathematical models are statistics concerning system primitives, such as the number of agents working, the rate at which calls arrive, the time required for a customer to be served, and the length of time customers are willing to wait on hold before they hang up the phone and *abandon* the queue. Outputs are performance measures, such as the distribution of time that customers wait "on hold" and the fraction of customers that abandon the queue before being served. In practice, the number of agents working becomes a control parameter which can be increased or decreased to attain the desired efficiency-quality tradeoff.

Often, estimates of two or more moments of the primitives are needed to calibrate queueing models, and in many cases, the models make distributional assumptions concerning the primitives. In theory, the data required to validate and properly tune these models should be readily available, since computers track and control the minutest details of every call's progress through the system. It is thus surprising that operational data, collected at an appropriate level of detail, has been scarcely available. The data that are typically collected and used in the call-center industry are simple *averages* that are calculated for the calls that arrive within fixed intervals of time, often 15 or 30 minutes. There is thus a lack of documented, comprehensive, empirical research on call-center performance that employs more detailed data.

The immediate goal of our study is to fill this gap. In this paper, we summarize a comprehensive analysis of operational data from a bank call center. The data span all twelve months of 1999 and are collected at the level of individual calls. Our data source consists of over 1,200,000 calls that arrived to the center over the year. Of these, about 750,000 calls terminated in an interactive voice response unit (IVR or VRU), a type of answering machine that allows customers serve themselves. The remaining 450,000 callers asked to be served by an agent, and we have a record of the event-history of each of these calls.

1.1 Highlights of results

1. The arrival times of calls requesting service by an agent are extremely well modelled as an inhomogeneous Poisson process. This is not a surprise. The process has a smoothly varying rate function, λ, that depends, in part, on the date, time of day, type and priority of the call. (See Section 4.1.)

2. The function λ is a hidden (or latent) rate function that it is not observed. That is, λ is not functionally determined by date, time, and call-type information. Because of this feature, we develop models to predict statistically the number of arrivals as a function of only these repetitive features. The methods also allows us to construct confidence bounds related to these predictions. See also Brown, Mandelbaum, Sakov, Shen, Zeltyn and Zhao (2002), where these models are more fully developed. (See Sections 4.2 and 7.4.)

3. The times required to serve customers of various types have lognormal distributions. This is quite different from the exponential distribution customarily assumed for such times, a finding that has potentially important implications for the queueing theory used to model the process. (See Section 5.3.)

4. The lognormal nature of service times has also led us to develop new methods for nonparametric estimation of regression models involving lognormal errors, as well as for the generation of confidence and prediction intervals. (The mean time, as opposed to the median or the mean of the log-time, is of central interest to queueing theory, and this necessitates more precise tools for statistical inference about the mean.) Some of these new methods are reported here; others will be included in Shen (2002). (See Section 5.4.)

5. A surprising feature is the presence of a disproportionate number of extremely short service times, under 5 seconds. (Presumably these are mainly due to agents who prematurely disconnect from the customer without offering any real service.) This reflects a behavioral feature of service agents that needs to be taken into account (and discouraged by various means) in engineering a queueing system. This feature also needs to be suitably accommodated in order to correctly estimate the parameters and features of the service-time distribution. (See Section 5.1.)

6. A joint feature of the arrivals and service times is that they are noticeably positively correlated as a function of time-of-day (and given customer types and priorities). Thus, during the busiest times of day (10am-12pm and 2pm-4pm) the service times are also, on average, longest. We are not sure what the explanation is for this and have developed several alternate hypotheses. Section 7 reports an examination of the data to see whether one of these can be identified as the primary cause. Determination of the cause could noticeably affect calculations of the so-called "offered load" that is central to queueing-theoretic calculation of system performance. (See Sections 4.1, 5.4 and 7.2.)

7. The hazard rate of customer patience is estimated using a competing-risks model for censored data. The very large size of our data set makes this an unusual situation for such an analysis, and it enables us to develop new nonparametric methods and graphical techniques. One notably finding is a marked short-run increase in the propensity of customers to abandon the queue at times shortly after they receive a message that asks them to be patient. We also examine two different patience indices that measure the relationship between customer patience and waiting time, and investigate the observed relationship between them. (See Sections 6.3 and 6.4.)

8. We examine the validity of a queueing-theoretic rule that relates the average waiting time in queue to the probability of abandonment. This "ratio" rule is derived in Baccelli and Hebuterne (1981) and Zohar, Mandelbaum and Shimkin (2002) and holds for models with exponentially distributed patience. We find that the rule seems to hold with reasonable precision for our data, even though customers' patience distribution is noticeably non-exponential. We are currently searching for a theoretical explanation of this empirical observation. (See Section 8.1.)

9. It is also possible to check the applicability of a multiple-server generalization of the classical Khintchine-Pollaczek formula from queueing theory (Iglehart and Whitt, 1970; Whitt, 2002). The data show that this formula does not provide a reasonable model for the call center under study, a result that is not surprising in that the formula ignores customer abandonment behavior. Conversely, alternatives that explicitly account for abandonment, such as Baccelli and Hebuterne (1981) and Garnett, Mandelbaum and Reiman (2002), perform noticeably better. (See Section 8.)

10. It is of interest to note that relatively simple queueing models can turn out to be surprisingly robust. For example, when carefully tuned, the so-called M/M/N+M model can be made to fit performance measures reasonably accurately, even though characteristics of the theoretical primitives clearly do not conform with the empirical ones. (See Section 8.2.)

Our study also constitutes a prototype that paves the way for larger-scale studies of call centers. Such a study is now being conducted under the auspices of the Wharton Financial Institutions Center.

This paper is part of a larger effort to use both theoretical and empirical tools to better characterize call center operations and performance. Mandelbaum, Sakov and Zeltyn (2000) presents a comprehensive descriptive analysis of our call-by-call database. Zohar et al. (2002) considers customer patience and abandonment behavior. Gans, Koole and Mandelbaum (2002) reviews queueing and related capacity-planning models of call centers, and it describes additional sources of call-center data (marketing, surveys, benchmarking). Mandelbaum (2001) is a bibliography of more than 200 academic papers related to call-centers.

1.2 Structure of the paper

The paper is structured as follows. In Section 2, we provide theoretical background on queueing models of call centers. Next, in Section 3, we describe the call center under study and its database.

Each of Sections 4 to 6 is dedicated to the statistical analysis of one of the stochastic primitives of the queueing system: Section 4 addresses call arrivals; Section 5, service durations; and Section 6, tele-queueing and customer patience. Section 6 also analyzes customer waiting times, a performance measure that, interestingly, is deeply intertwined with the abandonment primitive.

A synthesis of the primitive building blocks is typically needed for operational understanding. To this end, Section 7 discusses prediction of the arriving "workload", which is essential in practice for setting suitable service staffing levels.

Once each of the primitives has been analyzed, one can also attempt to use existing queueing theory, or modifications thereof, to describe certain features of the holistic behavior of the system. In Section 8 we conclude with analyses of this type. We validate some classical theoretical results from queueing theory and refute others.

2 QUEUEING MODELS OF CALL CENTERS

Call centers agents provide *tele-services*. As they speak with customers over the phone, they interact with a computer terminal, inputting and retrieving information related to customers and their requests. Customers, who are only virtually present, are either being served or are waiting in what we call a *tele-queue*, a phantom queue which they share, invisible to each other and to the agents who serve them. Customers wait in this queue until one of two things happen: an agent is allocated to serve them (through supporting software), or they become *impatient* and *abandon* the tele-queue.

Queueing theory was conceived by A.K. Erlang (Erlang, 1911; Erlang, 1917) at the beginning of the 20th century and has flourished since to become one of the central research themes of Operations Research (For example, see Wolff 1989, Buzacott and Shanthikumar 1993 and Whitt 2002). In a queueing model of a call center, the customers are callers, the servers are telephone agents or communications equipment, and queues are populated by callers that await service.

The queueing model that is simplest and most widely used in call centers is the so-called M/M/N system, sometimes called the Erlang-C model.[1] Given arrival rate λ, average service duration μ^{-1}

[1]The first M in "M/M/N" comes from Poisson arrivals, equivalently exponentially distributed, or Markovian, interarrival times. The second M, from Markovian (exponential) service times. The N denotes the number of servers working in parallel.

and N servers working in parallel, the Erlang-C formula, $C(\lambda, \mu, N)$, describes, theoretically, the long-run fraction of time that all N servers will be simultaneously busy, or (via PASTA, see Wolff 1989) the fraction of customers who are delayed in queue before being served. In turn, it allows for calculation of the theoretical distribution of the length of time an arriving customer will have to wait in queue before being served.

The Erlang-C model is quite restrictive, however. It assumes, among other things, a steady-state environment in which arrivals conform to a Poisson process, service durations are exponentially distributed, and customers and servers are statistically identical and act independently of each other. It does not acknowledge, among other things, customer abandonment behavior, time-dependent parameters, or customer heterogeneity. We refer the reader to Sze (1984), Harris, Hoffman and Saunders (1987), Garnett and Mandelbaum (1999), Garnett et al. (2002) and Zohar et al. (2002) for an elaboration of the Erlang-C's shortcomings. An essential task of queueing theorists is to develop models that account for the most important of these effects.

Queueing science seeks to determine which of these effects is most important. For example, Garnett et al. (2002) develops both exact and approximate expressions for "Erlang-A" systems, which explicitly model customer patience (time to abandonment) as exponentially distributed. Data analysis of the arrival process, service times, and customer patience can confirm or refute whether system primitives conform to the Erlang-C and A models' assumptions. Empirical analysis of waiting times can help us to judge how well the two models predict customer delays – whether or not their underlying assumptions are met.

3 THE CALL CENTER OF BANK ANONYMOUS

The source of our data is a small call center of one of Israel's banks. (The small size has proved very convenient, in many respects, for a pioneering field study.) The center provides several types of services: information for current and prospective customers, transactions for checking and savings accounts, stock trading, and technical support for users of the bank's internet site. On weekdays (Sunday through Thursday in Israel) the center is open from 7am to midnight; it closes at 2pm on Friday, for the weekend, and reopens around 8pm on Saturday. During working hours, at most 13 regular agents, 5 internet agents, and one shift supervisor may be working.

A simplified description of the path each call follows through the center is as follows. A customer calls one of several telephone numbers associated with the call center, the number depending on the type of service sought. Except for rare busy signals, the customer is then connected to a VRU and identifies herself. While using the VRU, the customer receives recorded information, general and customized (e.g. an account balance). It is also possible for the customer to perform some self-service transactions here, and 65% of the bank's customers actually complete their service via

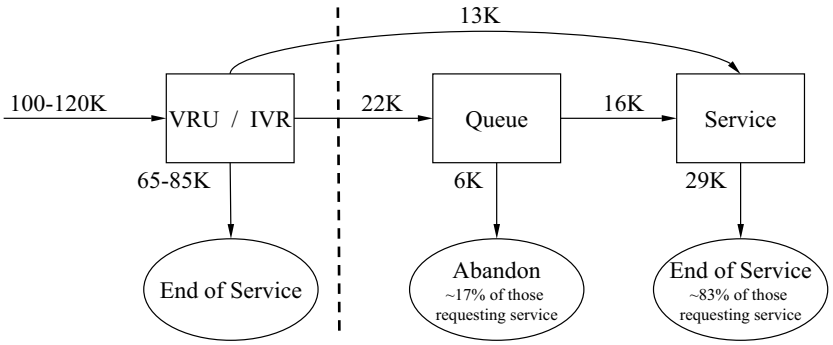
the VRU. The other 35% indicate the need to speak with an agent. If there is an agent free who is capable of performing the desired service, the customer and the agent are matched to start service immediately. Otherwise the customer joins the tele-queue.

Customers in the tele-queue are nominally served on a first-come first-served (FCFS) basis, and customers' places in queue are distinguished by the time at which they arrive to the queue. In practice, the call center operates a priority system with two priorities - high and low - and moves high-priority customers up in queue by subtracting 1.5 minutes from their actual arrival times. Mandelbaum et al. (2000) compares differences between the behaviors of the two groups of customers.

While waiting, each customer periodically receives information on her progress in the queue. More specifically, she is told the amount of time that the first in queue has been waiting, as well as her approximate location in the queue. The announcement is replayed every 60 seconds or so, with music, news, or commercials intertwined.

Figure 1 provides a schematic summary of the event history (process flow) of calls through the system. In the figure, the numbers next to arrows represent approximate numbers of calls each month that arrived to the VRU, queue, abandoned, were served. From the figure, we see that in each of the 12 months of 1999, roughly 100,000-120,000 calls arrived to the system with 65,000-85,000 terminating in the VRU. The remaining 30,000-40,000 calls per month involved callers who exited the VRU indicating a desire to speak to an agent. The percentages within the ovals show that about 80% of those requesting service were, in fact served, and about 20% abandoned before being served. The focus of our study is the set of 30,000-40,000 calls each month that crossed Figure 1's dashed line and queued or were immediately served.

Figure 1: Event history of calls (units are calls per month)



Each call that crosses the dashed line can be thought of as passing through up to three stages, each of which generates distinct data. The first is the *arrival* stage, which is triggered by the call's exit from the VRU and generates a record of an *arrival time*. If no appropriate server is available, then the call enters the *queueing* stage. Three pieces of data are recorded for each call that queues: the time it entered the queue; the time it exited the queue; and the manner in which it exited the

queue, by being served or abandoning. The last stage is *service*, and data that are recorded are the starting and ending times of the service. Note that calls that are served immediately skip the queueing stage, and calls that abandon never enter the service stage.

In addition to these time stamps, each call record in our database includes a categorical description of the type of service requested. The main call types are Regular (PS in the database), Stock Transaction (NE), New Customer (NW), and Internet Assistance (IN). (Two other types of call – Service in English (PE) and Outgoing Call (TT) – exist. Together, they accounted for less than 2% of the calls in the database, and they have been omitted from the analysis reported here.)

Our database includes calls for each month of 1999, and over the year there were two operational changes that are important to note. First, from January through July, all calls were served by the same group of agents, but beginning in August, internet (IN) customers were served by a separate pool of agents. Thus, from August through December the center can be considered to be two separate service systems, one for IN customers and another for all other types. Second, as will be noted in Section 5, one aspect of the service-time data changed at the end of October. In several instances this paper's analyses are based on only the November and December data. In other instances we have used data from August through December. Given the changes noted above, this ensures consistency throughout the manuscript. November and December were also convenient because they contained no Israeli holidays. In these analyses, we also restrict the data to include only regular weekdays – Sunday through Thursday, 7am to midnight – since these are the hours of full operation of the center. We have performed similar analyses for other parts of the data, and in most respects the November–December results do not differ noticeably from those based on data from other months of the year.

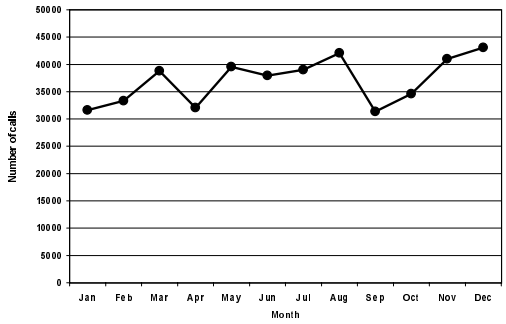
4 THE ARRIVAL PROCESS

Buffa, Cosgrove and Luce (1976) describes four levels of representation of a call center that differ according to their time scales. Information about the system at these levels is required to support the complex task of efficiently staffing a center. The two top levels are the monthly ("strategic") level and daily ("tactical") level. Figure 2 shows plots of the arrival of calls at these two levels.

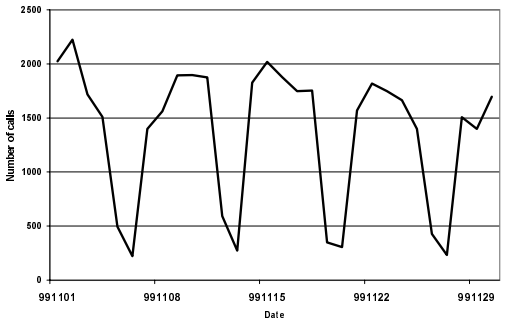
The strategic level is presented in an annual plot with month being the time unit. The decrease in calls in April and September is due to many holidays in these months. Strategic information is required to determine how many agents are needed in all, perhaps by season, and it affects hiring and training schedules.

The tactical level is presented here in a monthly plot with day as the unit. The valleys occur during weekends when the center is only open for a few hours per day. Daily information is used to determine work assignments: given the total number of employees available, more or fewer agents

Figure 2: Arrival process plots (monthly and daily level)



Strategic level
(calls per month in 1999)



Tactical level
(calls per day in November)

are required to work each day, depending (in part) on the total number of calls to be answered.

It should be clear that information at both of these levels is important for proper operation of the center. We will concentrate our analysis on the finer levels, however: hourly ("operational"); and real time ("stochastic"). Queueing models operate at these levels, and they are also the levels that present the most interesting statistical challenges.

Figures 3 and 4 show, as a function of time of day, the average rate per hour at which calls come out of the VRU. These are composite plots for weekday calls in November and December. The plots show calls according to the major call types. The volume of Regular (PS) calls is much greater than that of the other 3 types; hence those calls are shown on a separate plot. These plots are kernel estimates using normal kernels. (The kernels have sd = 0.15 for the first plot and sd = 0.25 for the others. The bandwidths were visually chosen to preserve most of the regular variability evident in a histogram with 10 minute bin-widths.) For a more precise study of these arrival rates, including confidence and prediction intervals see Section 7 and also Brown et al. (2002) and Brown, Zhang and Zhao (2001).

Note the bimodal pattern of Regular call-arrival times in Figure 3. It is especially interesting that Internet service calls (IN) do not show a similar bimodal pattern and, in fact, have a peak in volume after 10pm. (We speculate that this peak can be partially explained by the fact that internet customers are sensitive to telephone rates, which significantly decrease in Israel after 10pm, and they also tend to be late-night types of people.)

Figure 3: Arrivals (to queue or service); Regular Calls/Hr; Weekdays Nov.–Dec.

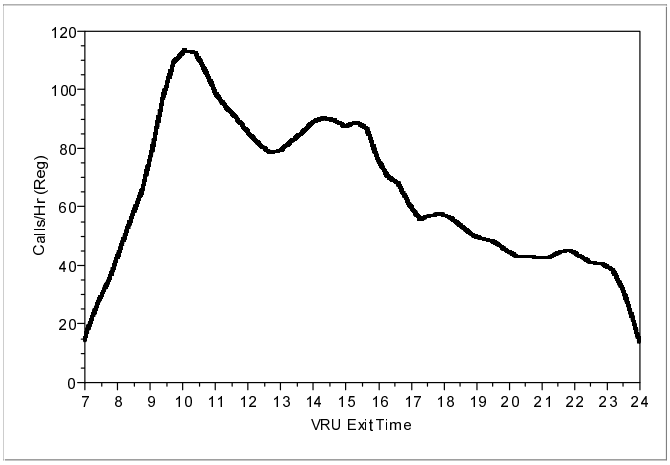
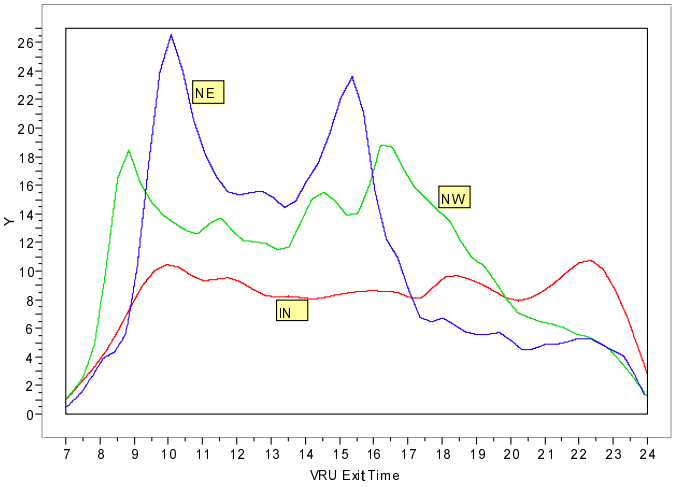


Figure 4: Arrivals (to queue or service); IN, NW, and NE Calls/Hr; Weekdays Nov.–Dec.



4.1 Arrivals are inhomogeneous Poisson

Classical theoretical models posit that arrivals form a Poisson process. It is well known that such a process results from the following behavior: there exist many potential, statistically identical callers to the call center; there is a very small yet non-negligible probability for each of them calling at any given minute, say, so that the average number of calls arriving within a minute is moderate; and callers decide whether or not to call independently of each other.

Common call-center practice assumes that the arrival process is Poisson with a rate that remains constant for blocks of time, often individual half-hours or hours. A call center will then fit a separate

queueing model for each block of time, and estimated performance measures may shift abruptly from one interval to the next.

A more natural model for capturing both stochastic and operational levels of detail is a time inhomogeneous Poisson process. Such a process is the result of time-varying probabilities that individual customers call, and it is completely characterized by its arrival-rate function. To be useful for constructing a (time inhomogeneous) queueing model, this arrival-rate function should vary smoothly and not too rapidly throughout the day. Smooth variation of this sort is familiar in both theory and practice in a wide variety of contexts, and seems reasonable here.

We now construct a test of the null hypothesis that arrivals of given types of calls form an inhomogeneous Poisson process, with the arrival rate varying slowly. This test procedure does not assume that the arrival rates (of a given type) depend only on the time-of-day and are otherwise the same from day to day. It also does not require the use of additional covariates to (attempt to) estimate the arrival rate for given date and time-of-day.

The first step in construction of the test involves breaking up the interval of a day into relatively short blocks of time. For convenience we used blocks of equal time-length, L, resulting in a total of I blocks, though this equality assumption could be relaxed. For the Regular (PS) data we used $L = 6$ minutes. For the other types we used $L = 60$ minutes, since these types involved much lower arrival rates. One can then consider the arrivals within a subset of blocks – for example, blocks at the same time on various days or successive blocks on a given day. Let T_{ij} denote the j-th ordered arrival time in the i-th block, $i = 1, \ldots, I$. Thus $T_{i1} \leq \ldots \leq T_{iJ(i)}$, where $J(i)$ denotes the total number of arrivals in the i-th block. Then define $T_{i0} = 0$ and

$$R_{ij} = (J(i) + 1 - j) \left(- \log \left(\frac{L - T_{ij}}{L - T_{i,j-1}} \right) \right), \quad j = 1, ..., J(i).$$

Under the null hypothesis that the arrival rate is constant within each given time interval, the $\{R_{ij}\}$ will be independent standard exponential variables as we now discuss.

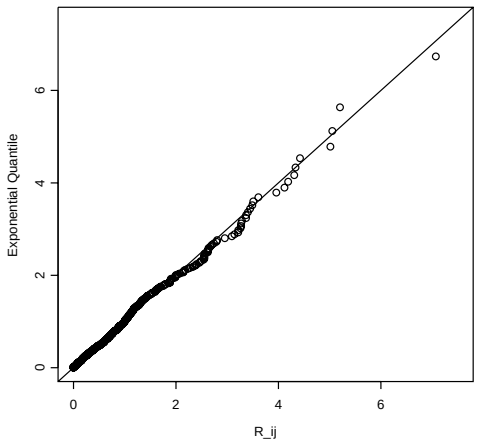
Let U_{ij} denote the j-th (unordered) arrival time in the i-th block. Then the assumed constant Poisson arrival rate within this block implies that, conditionally on $J(i)$, the unordered arrival times are independent and uniformly distributed, i.e. $U_{ij} \overset{i.i.d.}{\sim} \mathrm{U}(0, L)$. Note that $T_{ij} = U_{i(j)}$. It follows that $\frac{L - T_{ij}}{L - T_{i,j-1}}$ are independent Beta$(J(i)+1-j, 1)$ variables. (See, for example, Problem 6.14.33(iii) in Lehmann 1986.) A standard change of variables then yields the exponentiality of the R_{ij}. (One may alternatively base the test on the variables $R^*_{ij} = j \left(- \log \frac{T_{ij}}{T_{i,j+1}} \right)$ where $j = 1, \ldots, J(i)$ and $T_{i,J(i)+1} = L$. Under the null hypothesis these will also be independent standard exponential variables.)

The null hypothesis does not involve an assumption that the arrival rates of different intervals are equal or have any other pre-specified relationship. Any customary test for the exponential distri-

bution can be applied to test the null hypothesis. For convenience we use the familiar Kolmogorov-Smirnov test, even though this may not have the greatest possible power against the alternatives of most interest.

Figures 5 and 6 display the results of two applications of this test. Figure 5 is an exponential quantile plot for the $\{R_{ij}\}$ computed from arrival times of the Regular (PS) calls arriving between 11:12am and 11:18am on all weekdays in November and December. Figure 6 is a similar plot from arrival of IN calls throughout Monday, November 23; from 7am to midnight. This was a typical midweek day in our data set.

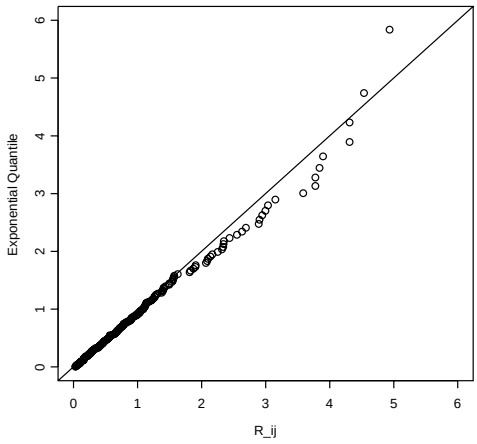
Figure 5: Exponential (λ=1) Quantile plot for $\{R_{ij}\}$ from Regular calls 11:12am – 11:18am; Nov. and Dec. weekdays



For both of the examples, pictured in Figures 5 and 6, the null hypothesis is not rejected, and we conclude that their data are consistent with the assumption of an inhomogeneous Poisson process for the arrival of calls. The respective Kolmogorov-Smirnov statistics have values K = 0.0316 (P-value = 0.2 with n = 420) and K = 0.0423 (P-value = 0.2 with n = 172). These results are typical of those we have obtained from various selections of blocks of the various types of calls. Thus, overall there is no evidence in this data set to reject a null hypothesis that the arrival of calls from the VRU is an inhomogeneous Poisson process.

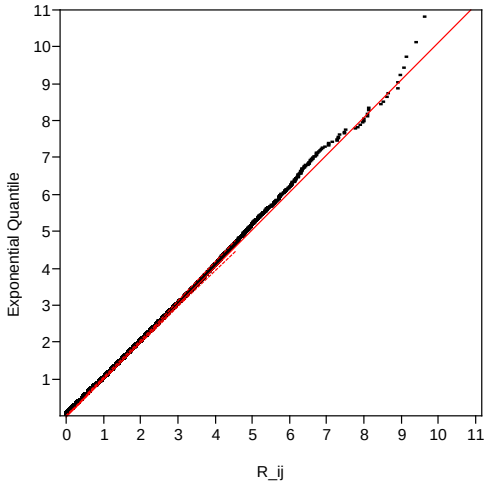
As a further demonstration that the arrival times follow an inhomogeneous Poisson process, we applied this method to the 48,963 Regular (PS) calls in November and December in blocks of 6 minutes for calls exiting the VRU on weekdays between 7am and 12pm. In view of the very large sample size we should expect the null hypothesis to be rejected, since the arrival rate is not exactly constant over 6 minute intervals, and also since the data are recorded in discrete units (of seconds) rather than being exactly continuous, as required by the ideal assumptions in the null hypothesis.

Figure 6: Exponential (λ=1) Quantile plot for $\{R_{ij}\}$ from Internet calls; Nov. 23



With this in mind, we view the fidelity of the exponential quantile plot in Figure 7 to a straight line as remarkable evidence that there is no discernible deviation of any practical importance from the assumption that the arrival data follow the law of an inhomogeneous Poisson process. (Two outlying values of R_{ij} have been omitted in Figure 7. They correspond to two 6-minute intervals in which the telephone switching system apparently malfunctioned.)

Figure 7: Exponential Quantile plot of $\{R_{ij}\}$ for all PS calls, weekdays in Nov.–Dec. (Two outliers omitted.)



We note that in this section, as well as in the following section, many tests are performed on the same data. For example, we test for the Poisson property over different blocks of time, different

days, different types of service etc. This introduces the problem of multiple testing: when many null hypotheses are rejected, a question should be asked if the rejections are real. One way to account for the problem is to use the FDR procedure (Benjamini and Hochberg, 1995). We do not use the procedure here since we are mainly illustrating what can be done with the data, rather than using them to make operational inferences. However, it is critical to use FDR or other procedures when making operational inferences.

4.2 The Poisson arrival-rates are not easily "predictable"

The inhomogeneous Poisson process described above provides a stochastic regularity that can sometimes be exploited. However, this regularity is most valuable if the arrival rates are known, or can be predicted on the basis of observable covariates. The current section examines the hypothesis that the Poisson rates can be written as a function of the available covariates: call type, time-of-day and day-of-the-week. If this were the case, then these covariates could be used to provide valid stochastic predictions for the numbers of arrivals. But, as we now show, this is not the case.

The null hypothesis to be tested is, therefore, that the Poisson arrival rate is a (possibly unknown) function $\lambda_{type}(d,t)$, where $type \in \{PS, NE, NW, IN\}$ may be any one of the types of customers, $d \in \{\text{Sunday}, \ldots, \text{Thursday}\}$ is the day of the week, and $t \in [7, 24]$ is the time of day. For this discussion, let Δ be a specified calendar date (e.g. November 7th), and let $N_{type,\Delta}$ denote the observed number of calls requesting service of the given type on the specified date. Then, under the null hypothesis, the $N_{type,\Delta}$ are independent Poisson variables with respective parameters

$$\mathrm{E}[N_{type,\Delta}] = \int_7^{24} \lambda_{type}(d(\Delta), t)dt\,, \tag{1}$$

where $d(\Delta)$ denotes the day-of-week corresponding to the given date.

Under the null hypothesis, each set of samples for a given type and day-of-week, $\{N_{type,\Delta} \,|\, d(\Delta) = d\}$, should consist of independent draws from a common Poisson random variable. If so, then one would expect the sample variance to be approximately the same as the sample mean. For example, see Agresti (1990) and Jongbloed and Koole (2001) for possible tests.

Table 1 gives some summary statistics for the observed values of $\{N_{NE,\Delta} \,|\, d(\Delta) = d\}$ for weekdays in November and December. Note that there were 8 Sundays and 9 each of Monday through Thursday during this period. A glance at the data suggests that the $\{N_{NE,\Delta} \,|\, d(\Delta) = d\}$ are not samples from Poisson distributions. For example, the sample mean for Sunday is 163.38, and the sample variance is 475.41. This observation can be validated by a formal test procedure, as described in the following paragraphs.

Brown and Zhao (2001) present a convenient test for fit to an assumption of independent Poisson

Table 1: Summary statistics for observed values of $N_{NE,\Delta}$, weekdays in Nov. and Dec.

Day-of-Week (d)	# of Dates (n_d)	Mean	Variance	Test Statistic (V_d)	P-value
Sunday	8	163.38	475.41	20.41	0.0047
Monday	9	188.78	1052.44	42.26	0.0000
Tuesday	9	199.67	904.00	38.64	0.0000
Wednesday	9	185.00	484.00	21.23	0.0066
Thursday	9	183.89	318.61	13.53	0.0947
ALL	44			$V_{all} = 136.07$	0.00005

variables. This is the test employed below. The background for this test is Anscombe's variance stabilizing transformation for the Poisson distribution (Anscombe, 1948).

To apply this test to the variables $\{N_{NE,\Delta} \,|\, d(\Delta) = d\}$, calculate the test statistic

$$V_d \;=\; 4 \;\times \sum_{\{\Delta \,|\, d(\Delta)=d\}} \left(\sqrt{N_{NE,\Delta} + 3/8} \;\; - \;\; \frac{1}{n_d} \sum_{\{\Delta \,|\, d(\Delta)=d\}} \sqrt{N_{NE,\Delta} + 3/8} \right)^2 ,$$

where n_d denotes the number of dates satisfying $d(\Delta) = d$. Under the null hypothesis that the $\{N_{NE,\Delta} \,|\, d(\Delta) = d\}$ are independent identically distributed Poisson variables, the statistic V_d has very nearly a Chi-squared distribution with $(n_d - 1)$ degrees of freedom. The null hypothesis should be rejected for large values of V_d. Table 1 gives the values of V_d for each value of d, along with the P-values for the respective tests. Note that for these five separate tests the null hypothesis is decisively rejected for all but the value $d =$ Thursday.

It is also possible to use the $\{V_d\}$ to construct a test of the pooled hypothesis that the $N_{NE,\Delta}$ are independent Poisson variables with parameters that depend only on $d(\Delta)$. This test uses $V_{all} = \sum V_d$. Under the null hypothesis this will have (very nearly) a Chi squared distribution with $\sum(n_d - 1)$ degrees of freedom, and the null hypothesis should be rejected for large values of V_{all}. The last row of Table 1 includes the value of V_{all}, and the P-value is less than or equal to 0.00005.

The qualitative pattern observed in Table 1 is fairly typical of those observed for various types of calls, over various periods of time. For example, a similar set of tests for type NW for November and December yields one non-significant P-value (P = 0.2 for $d(\Delta) =$ Sunday), and the remaining P-values are vanishingly small. A similar test for type PS (Regular) yields all vanishingly small P-values.

The tests above can also be used on time spans other than full days. For example, we have constructed similar tests for PS calls between 7am and 8am on weekdays in November and December.

(A rationale for such an investigation would be a theory that early morning calls – before 8am – arrive in a more predictable fashion than those later in the day.) All of the test statistics are extremely highly significant: for example the value of V_{all} is 143 on 39 degrees of freedom. Again, the P-value is less than or equal to 0.00005.

In summary, we saw in Section 4.1 that, for a given customer type, arrivals are inhomogeneous Poisson, with rates that depend on time of day as well as on other possible covariates. In Section 4.2 an attempt was made to characterize the exact form of this dependence, but ultimately the hypothesis was rejected that the Poisson rate was a function only of these covariates. For the operation of the call center it is desirable to have predictions, along with confidence bands, for this rate. We return to this issue in Section 7.

5 SERVICE TIME

The last phase in a successful visit to the call center is typically the service itself. Table 2 summarizes the mean, SD and median service times for the four types of service of main interest. The very few calls with service time larger than an hour were not considered (i.e. we treat them as outliers). Adding these calls has little effect on the numbers.

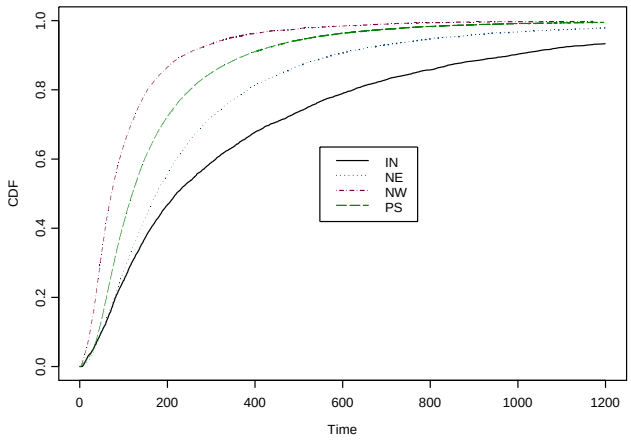
Table 2: Service time by type of service, truncated at 1 hour, Nov.–Dec.

	Overall	Regular Service (PS)	New Customers (NW)	Internet Consulting (IN)	Stock Trading (NE)
Mean	201	179	115	401	270
SD	248	189	146	473	303
Med	124	121	73	221	175

Internet consulting calls have the longest service times, and trading services are next in duration. New customers have the shortest service time (which is consistent with the nature of these calls). An important implication is that the workload that Internet consultation imposes on the system is more than its share in terms of percent of calls. This is an operationally important observation to which we will return in Section 7.

Figure 8 plots the cumulative distribution functions of service times by type. Note the clear stochastic ordering among the types, which strengthens the previously-discovered inequalities among mean service times.

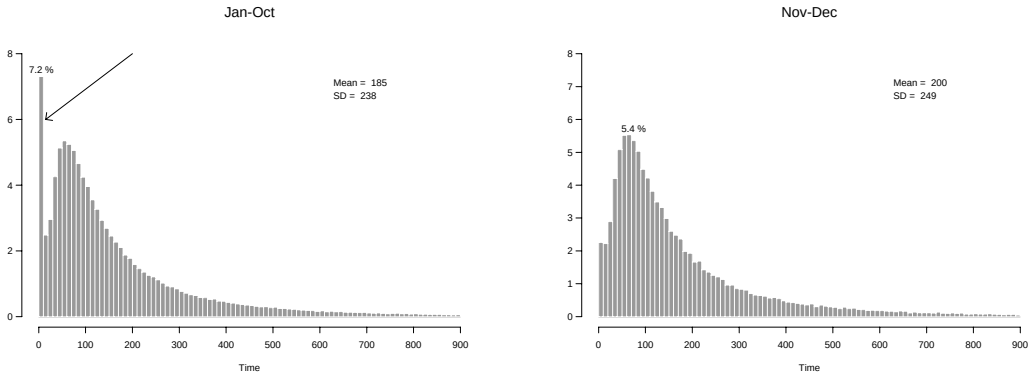
Figure 8: Service times cumulative distribution functions, by types, Nov.–Dec.



5.1 Very short service times

Figure 9 shows histograms of the combined service times for all types of service for January through October and for November–December. These plots resemble those for Regular Service calls alone, since the clear majority of calls are for regular service. Looking closely at the histograms, we see that, in the first 10 months of the year, the percent of calls with service shorter than 10 seconds was larger than the percent at the end of the year (7% vs. 2%).

Figure 9: Distribution of service time



Service times shorter than 10 seconds are questionable. And, indeed, the manager of the call center discovered that short service times were primarily caused by agents that simply *hung-up* on customers to obtain extra rest-time. (The phenomenon of agents "abandoning" customers is not

uncommon; it is often due to distorted incentive schemes, especially those that over-emphasize short average talk-time, or equivalently, the total number of calls handled by an agent). The problem was identified and steps were taken to correct it in October of 1999, after a large number of customers had complained about being disconnected. For this reason, in the later analysis we focus on data from November and December. Suitable analyses can be constructed for the entire year through the use of a mixture model or (somewhat less satisfactorily) by deleting from the service-time analysis all calls with service times under 10 seconds.

5.2 On service times and queueing theory

Most applications of queueing theory to call centers assume exponentially distributed service times as their default. The main reason is the lack of empirical evidence to the contrary, which leads one to favor convenience. Indeed, models with exponential service times are amenable to analysis, especially when combined with the assumption that arrival processes "are" Poisson processes (a rather natural one for call centers, see Section 4.1). The prevalent M/M/N (Erlang-C) model is an example.

In more general queueing formulae, the service time often affects performance measures through its squared-coefficient-of-variation $C^2 = \sigma^2/\mathrm{E}^2$, E being the average service time, and σ its standard deviation. For example, a useful approximation for the average waiting time in an M/G/N model (Markovian arrivals, Generally distributed service times, N servers), is given by Whitt (1993):

$$\mathrm{E}[\text{Wait for M/G/N}] \approx \mathrm{E}[\text{Wait for M/M/N}] \ \times \ \frac{(1+C^2)}{2}.$$

Thus, average wait with general service times is multiplied by a factor of $(1+C^2)/2$ relative to the wait under exponential service times. For example, if service times are, in fact, exponential then the factor is 1, as should be; deterministic service times *halve* the average wait of exponential; and finally, based on Table 2, our empirical $(1+C^2)/2 = 1.26$.

In fact, in the approximation above and many of its "relatives", service times manifest themselves only through their means and standard deviations. Consequently, for practical purposes, if means and standard deviations are close to each other, then one assumes that system performance will be close to that with exponential service times.

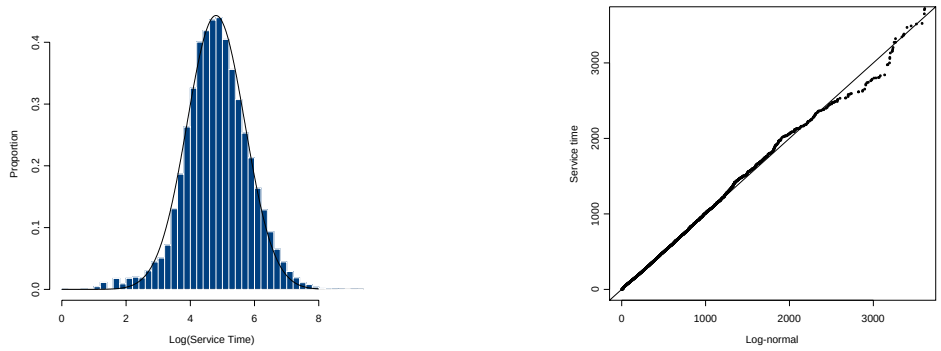
However, for large call centers with high levels of agent utilization, this assumption may not hold: simulation studies indicate that the *entire* distribution of service time may become very significant (for example, see Mandelbaum and Schwartz 2002). While ours is a small call center with moderate utilization levels, its service-time distributions are, nevertheless, likely to be similar to those of larger, more heavily utilized systems. Thus, it is worthwhile investigating the distribution of its service times.

5.3 Service times are lognormal

Looking at Figure 9, we see that the distribution of service times is clearly not exponential, as assumed by standard queueing theory. In fact, after separating the calls with very short service times, our analysis reveals a remarkable fit to the *lognormal* distribution.

The left panel of Figure 10 shows the histogram of log(service time) for November and December, in which the short service phenomenon was absent or minimal. We also superimpose the best fitted normal density as provided by Brown and Hwang (1993). The right panel shows the lognormal Q-Q plot of service time, which does an amazingly good imitation of a straight line. Both plots suggest that the distribution of Service Time is very nearly lognormal. We only provide the graphs to qualitatively support our claim of lognormality. The reason is that, given the large sample size we have, the Kolmogorov-Smirnov test (or any other goodness-of-fit test) will always reject the null hypothesis of lognormality due to a small deviation from the lognormal distribution.

Figure 10: Histogram, QQ Plot of Log(Service Time) (Nov.–Dec.)



After excluding short service times, the strong resemblance to a lognormal distribution also holds for all other months. It also holds for various types of callers even though the parameters depend on the type of call. This means that, in this case, a mixture of log-normals is log-normal empirically, even though mathematically this cannot hold. The same phenomenon occurs in Section 6.1, when discussing the exponential distribution. We refer the reader to Mandelbaum et al. (2000) where the phenomenon is discussed in the context of the exponential distribution.

Lognormality of processing times has been occasionally recognized by researchers in telecommunications and psychology. Bolotin (1994) shows empirical results which suggest that the distribution of the logarithm of call duration is normal for individual telephone customers and a mixture of normals for "subscriber-line" groups. Ulrich and Miller (1993) and Breukelen (1995) provide theoretical arguments for the lognormality of reaction times using models from mathematical psychology. Man-

delbaum and Schwartz (2002) use simulations to study the effect of lognormally distributed service times on queueing delays.

In our data, lognormality appears to also hold quite well at lower levels: for all types and priorities of customers, for individual servers, for different days of the week, and also when conditioning on time of day, as in Section 5.4. Brown and Shen (2002) gives a more detailed analysis of service times.

5.4 Regression of log service times on time-of-day

The important implication of the excellent fit to a lognormal distribution is that we can apply standard techniques to regress log(service time) on various covariates, such as time-of-day. For example, to model the mean service time across time-of-day, we can first model the mean and variance of the log(service time) across time-of-day and then transform the result back to the service-time scale. (Shen (2002) contains a detailed analysis of service times against other covariates, such as the identities of individual agents (servers), as well as references to other literature involving lognormal variates.)

Let S be a lognormally distributed random variable with mean ν and variance τ^2, then $Y = \log(S)$ will be a normal random variable with some mean μ and variance σ^2. It is well known that

$$\nu = e^{\mu + \frac{1}{2}\sigma^2}. \tag{2}$$

We use (2) as a basis for our proposed methodology. Suppose we wish to estimate $\nu = E(S)$ and construct an associated confidence interval. If we can derive estimates for μ and σ^2, then $\hat{\nu} = e^{\hat{\mu} + \frac{\hat{\sigma}^2}{2}}$ will be a natural estimate for ν according to (2). Furthermore, in order to provide a confidence interval for ν, we need to derive confidence intervals for μ and σ^2, or more precisely, for $\mu + \sigma^2/2$.

For our call center data, let S be the service time of a call and T be the corresponding time-of-day that the call begins service. Let $\{S_i, T_i\}_{i=1}^n$ be a random sample of size n from the joint distribution of $\{S, T\}$ and sorted according to T_i. Then $Y_i = \log(S_i)$ will be the Log(Service Time) of the calls, and these are (approximately) normally distributed, conditional on T_i. We can fit a regression model of Y_i on T_i as

$$Y_i = \mu(T_i) + \sigma(T_i)\epsilon_i, \tag{3}$$

where $\epsilon_i | T_i$ are i.i.d. $N(0, 1)$.

5.4.1 Estimation of $\mu(\cdot)$

If we assume that $\mu(\cdot)$ has a continuous third derivative, then we can use local quadratic regression to derive an estimate for $\mu(\cdot)$. See Loader (1999). Suppose $\hat{\mu}(t_0)$ is an local quadratic estimate for

$\mu(t_0)$, then an approximate $100(1-\alpha)\%$ confidence interval for $\mu(t_0)$ is

$$\hat{\mu}(t_0) \pm z_{\alpha/2}\text{se}_\mu(t_0), \tag{4}$$

where $\text{se}_\mu(t_0)$ is the standard error of the estimate of the mean at t_0 from the local quadratic fit.

5.4.2 Estimation of $\sigma^2(\cdot)$

Our estimation of the variance function $\sigma^2(\cdot)$ is a two-step procedure. At the first step, we regroup the observations $\{T_i, Y_i\}_{i=1}^n$ into consecutive non-overlapping pairs $\{T_{2i-1}, Y_{2i-1}; T_{2i}, Y_{2i}\}_{i=1}^{\lfloor n/2 \rfloor}$. The variance at T_{2i}, $\sigma^2(T_{2i})$, is estimated by a squared pseudo-residual, D_{2i}, of the form $(Y_{2i-1}-Y_{2i})^2/2$, a so-called difference-based estimate. The difference-based estimator we use here is a simple one that suffices for our purposes. In particular, our method yields suitable confidence intervals for the estimation of σ^2. More efficient estimators might slightly improve our results. There are many other difference-based estimators in the literature. See Müller and Stadtmüller (1987), Hall, Kay and Titterington (1990), Dette, Munk and Wagner (1998) and Levins (2002) for more discussions of difference-based estimators.

During the second step, we treat $\{T_{2i}, D_{2i}\}_{i=1}^{\lfloor n/2 \rfloor}$ as our observed data points and apply local quadratic regression to obtain $\hat{\sigma}^2(t_0)$. Part of our justification is that, under our model (3), the $\{D_{2i}\}$'s are (conditionally) independent given the $\{T_{2i}\}$'s. Similar to (4), a $100(1-\alpha)\%$ confidence interval for $\sigma^2(t_0)$ is approximately

$$\hat{\sigma}^2(t_0) \pm z_{\alpha/2}\text{se}_{\sigma^2}(t_0).$$

Note that we use $z_{\alpha/2}$ as the cutoff value when deriving the above confidence interval, rather than a quantile from a Chi-square distribution. Given our large data set the degree of freedom is large, and a Chi-square distribution can be approximated well by a normal distribution.

5.4.3 Estimation of $\nu(\cdot)$

We now use $\hat{\mu}(t_0)$ and $\hat{\sigma}^2(t_0)$ to estimate $\nu(t_0)$, as $e^{\hat{\mu}(t_0)+\hat{\sigma}^2(t_0)/2}$. Given the estimation methods used for $\mu(t_0)$ and $\sigma^2(t_0)$, $\hat{\mu}(t_0)$ and $\hat{\sigma}^2(t_0)$ are asymptotically independent, which gives us

$$\text{se}(\hat{\mu}(t_0) + \hat{\sigma}^2(t_0)/2) \approx \sqrt{\text{se}_\mu(t_0)^2 + \text{se}_{\sigma^2}(t_0)^2/4}.$$

When the sample size is large, we can assume that $\mu(\cdot) + \sigma^2(\cdot)/2$ has an approximately normal distribution. Then the corresponding $100(1-\alpha)\%$ confidence interval for $\nu(t_0)$ is

$$\exp\left((\hat{\mu}(t_0) + \hat{\sigma}^2(t_0)/2) \pm z_{\alpha/2}\sqrt{\text{se}_\mu(t_0)^2 + \text{se}_{\sigma^2}(t_0)^2/4}\right).$$

5.4.4 Application and model diagnostics

In the following analysis, we apply the above procedure to the weekday calls of November and December. The results for two interesting service types are shown in Figures 11 and 12, below. There are 42,613 Regular Service (PS) calls and 5,066 Internet Consulting (IN) calls. To produce the figures, we use the tricube function as the kernel and nearest-neighbor type bandwidths. The bandwidths are subjectively chosen to generate interesting curves that are nearly free of (apparently) extraneous wiggles.

Figure 11 shows the mean service time for PS calls as a function of time-of-day, with 95% confidence bands. Note the prominent bimodal pattern of mean service time across the day for PS calls. The accompanying confidence band shows that the changing pattern is highly significant. Average service times are longest around 10:00am and 3:00pm. The changing pattern of the overall average (across all types of calls) will be similar, because about 70% of the calls are Regular Service calls.

Also notice that the pattern nicely resembles that for *arrival rates* of PS calls. (See Figure 3 in Section 4.) Call center managers should take this phenomenon into account when making operational decisions. We will come back to this issue and explore possible explanations when we analyze system workload in Section 7.

Figure 11: Mean Service Time (PS) versus Time-of-day (95% CI)

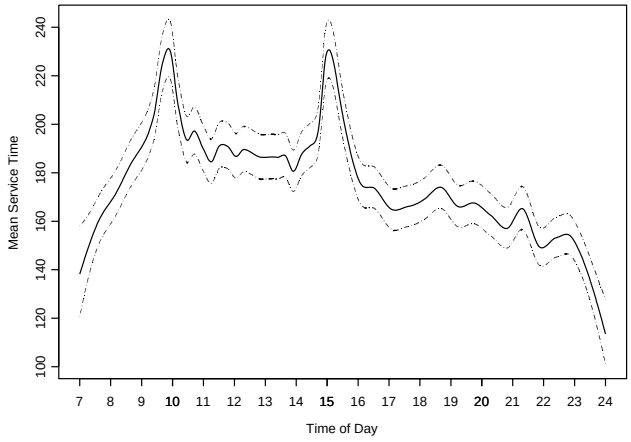
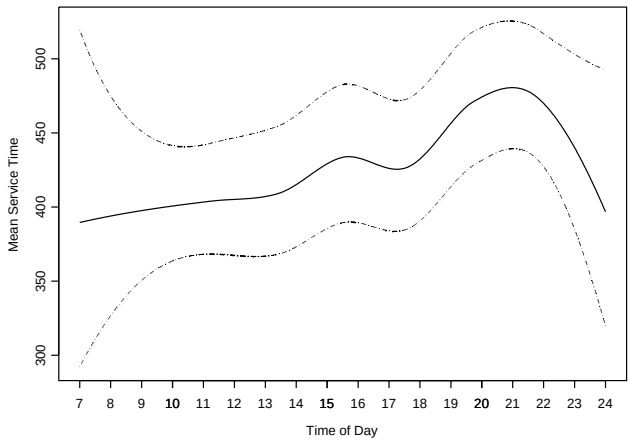


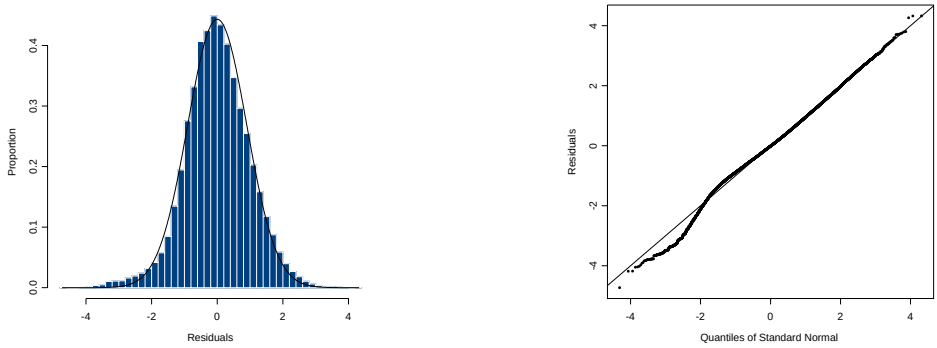
Figure 12 plots an analogous confidence band for IN calls. One interesting observation is that IN calls do not show a similar bimodal pattern. We do see some fluctuations during the day, but they are only mildly significant, given the wide confidence band. Also notice that the entire confidence band for IN calls lies above that of PS calls. This reflects the same stochastic dominance that was observed in Figure 8.

Figure 12: Mean Service Time (IN) versus Time-of-day (95% CI)



We close this section with some diagnostics of our model, looking at residuals from the regression of log(service time) on time-of-day for PS calls. Figure 13's histogram and normal quantile plot of residuals show that the residuals are fairly normal, and they provide additional validation of the lognormality of the service times.

Figure 13: Histogram, QQ Plot of The Residuals from Modeling Mean Log(Service Time) on Time-of-day (PS)



6 QUEUEING TIME: WAITING TIME FOR SERVICE OR ABANDONING

In Sections 4 and 5 we characterized two primitives of queueing models: the arrival process and service times. In each case we were able to directly observe and analyze the primitive under investigation.

Ideally, we would next address the last system primitive, customer patience and abandonment behavior, before considering system outputs, such as waiting time. Abandonment behavior and waiting times are deeply intertwined, however. By definition, all customers who abandon the tele-queue have waited. Furthermore, the times at which customers who are served *would have* abandoned, had they not been served, are not observed. Thus, the characterization of patience and time to abandonment is based on censored data.

We make three important outcome-based distinctions. The first is the difference between queueing time and waiting time. We use the convention that the latter does not account for zero-waits. This measure is more relevant for managers, especially when considered jointly with the fraction of customers that did wait. A second, more fundamental, distinction is between the waiting times of customers who were served and of those who abandoned. A third distinction is between the time that a customer *needs* to wait before reaching an agent versus the time that a customer is *willing* to wait before abandoning the system. The former is referred to as *virtual waiting time*, since it amounts to the time that a (virtual) customer, equipped with infinite patience, would have waited until being served. We refer to the latter as *patience*. Both measures are obviously of great importance, but neither is directly observable, and hence both must be estimated.

6.1 Waiting times are exponentially distributed

A well known queueing-theoretic result is that, in heavily loaded systems (in which essentially all customers wait, so that queueing time equals waiting time), queueing time should be exponentially distributed. See Kingman (1962) for an early result and Whitt (2002) for a recent text. Although our system is not very heavily loaded, we find that the empirical waiting time distribution conforms to the theoretical prediction.

Table 3 summarizes the mean, SD and median waiting time for all calls, as well as for calls stratified by outcome (A – Abandoned; S – Served) and by type of service. The waiting-time data have some observations that we consider to be outliers: for example, those with queueing time larger than 5 hours. Therefore, we have truncated the waiting time at 15 minutes. This captures about 99% of the data.

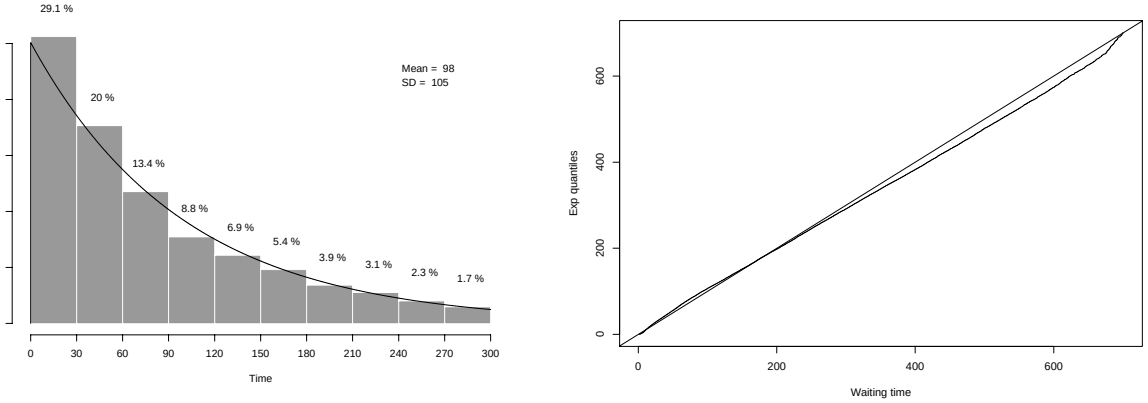
The left panel of Figure 14 shows the waiting-time histogram of all waiting calls. It resembles an

Table 3: Waiting time, truncated at 15 minutes (A – Abandoned; S – Served)

	Overall			PS		NE		NW		IN	
		A	S	A	S	A	S	A	S	A	S
Mean	98	78	105	62	96	99	114	88	136	140	159
SD	105	90	108	69	98	113	112	94	131	148	159
Med	62	51	67	43	62	55	78	58	92	86	103

exponential distribution, and the figure's right panel compares the waiting times to exponential quantiles, using a Q-Q plot. (The p-value for the Kolmogorov-Smirnov test for exponentiality is 0 however. This is not surprising in view of the large sample size of about 48,000.)

Figure 14: Distribution of waiting time (1999)



In fact, when restricted to customers reaching an agent, the histogram of waiting time resembles even more strongly an exponential distribution. Similarly, each of the means in Table 3 is close to the corresponding standard deviation, both for all calls and for those that reach an agent. This suggests (and was verified by QQ-plots) an exponential distribution also for each stratum, where a similar explanation holds: calls of type PS are about 70% of the calls. We also observe this exponentiality when looking at the waiting time stratified by months (Table 4).

6.2 Survival curves for virtual waiting time and patience

Both times to abandonment and times to service are censored data, and we apply survival analysis to help us estimate them. Denote by R the "patience" or "time willing to wait", by V the "virtual waiting time", and equip both with steady-state distributions. One actually samples

Table 4: Waiting time, truncated at 15 minutes

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Mean	68	76	119	109	96	85	105	114	84	73	101	111
SD	67	78	126	108	101	89	105	119	89	83	109	116
Med	46	50	75	72	62	55	69	72	54	45	63	71

$W = \min\{R, V\}$, as well as the indicator $1_{\{R<V\}}$ for observing R or V. To estimate the distribution of R, one considers all calls that reached an agent as censored observations, and vice versa for estimating the distribution of V. We make the assumption that (as random variables) R and V are independent given the covariates relevant to the individual customer. Under this assumption, the distributions of R and V (given the covariates) can be estimated using the standard Kaplan-Meier product-limit estimator. (See Section 6.5 for some cautionary remarks.)

In Figure 15, we plot the Kaplan-Meier estimates of the survival functions of R (time willing to wait), V (virtual waiting time) and $W = \min\{V, R\}$. A clear stochastic ordering emerges among the three distributions. Moreover, the same ordering arises at all months and across different types of service. The reason for the survival function of W being the lowest is obvious. In contrast, the stochastic ordering between V and R is interesting and informative. It indicates that customers are willing to wait (R) more than they need to wait (V), which suggests that our customers population consists of patient customers. Here we have implicitly, and only intuitively, defined the notion of a *patient customer* (to the best of our knowledge systematic research on this subject is lacking).

In Figure 16, we consider the survival functions of R for different types of service. Again, a clear stochastic ordering emerges. For example, we learn that customers performing stock trading (type 'NE') are willing to wait more then customers calling for regular services (type 'PS'). One might intuitively expect stock-trading customers to be less tolerant of delay. At the same time, their need for service, and their trust of the system to provide it, might be higher. Thus, a possible explanation for the ordering shown in Figure 16 is that type NE needs the service more urgently, and we are led to distinguish between tolerance for waiting and loyalty/persistency (which we do not pursue further).

Table 5 reports the means and standard deviations of the conventional Kaplan-Meier estimates for the distributions of V and R. The table is based on all weekday calls in November and December that waited in queue. We use a procedure that is conventional in several standard software packages. When the Kaplan-Meier estimate of the survival distribution of the event is defective (does not reach 0) this conventional estimator places all remaining survival probability at the largest event time.

Note that, for estimation involving R, the censoring rate is quite high. This results in several

Figure 15: Survival curves (Nov.–Dec.)

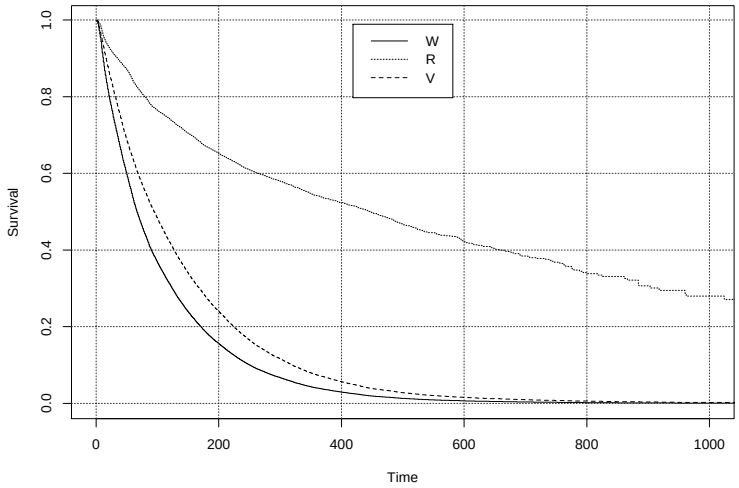
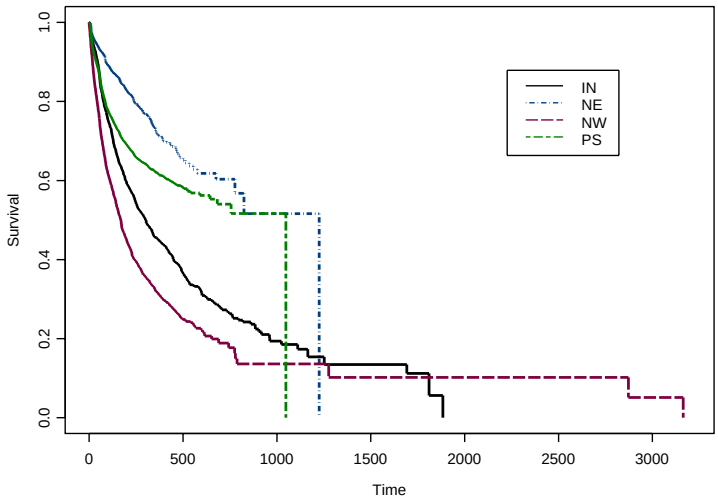


Figure 16: Survival curves for time willing to wait (Nov.–Dec.)



anomalies. In spite of the fact that one expects the true distributions to be skewed to the right, the estimated distributions are severely truncated. This is especially true for types PS and NE because they are more heavily censored. See Figure 16. A result of this is that the estimated means for PS and NE calls are much smaller than the estimated medians, while the opposite relation holds for

Table 5: Means, SDs and medians for V and R (Nov.–Dec.)

	Time willing to wait (R)			Time needs to wait (V)		
	Mean	SD	Median	Mean	SD	Median
All Combined	803	905	457	142	161	96
PS	642	446	1048	118	114	83
NE	806	471	1225	144	141	103
NW	535	885	169	227	251	193
IN	550	591	302	274	319	155

NW and IN calls. Also the estimated mean for the overall distribution (mean = 803) is much larger than that for its largest component, PS (mean = 642) and is approximately equal to the largest of the component means. We suspect that, as a consequence of this heavy censoring, the estimates for the mean of R are heavily biased (too small) and should be handled with care. Again this is especially true for types PS and NE in this table.

6.3 Hazard rates

Hazard rates are informative for understanding time-varying behavior. For example, local peaks in the hazard rates of the time willing to wait manifest a systematically increased tendency to abandon, while constant hazard rates indicate that the tendency to abandon remains the same, regardless of the past (memoryless). Palm (1953) was the first to describe impatience in terms of a hazard rate. He postulated that the hazard rate of the time-willing-to-wait is proportional to a customer's irritation due to waiting (thus defining, implicitly, the notion of irritation). Aalen and Gjessing (2001) advocate dynamic interpretation of the hazard rate, but warn against the possibility that the population hazard rate need not represent individual ones.

For this reason we have found it useful to construct nonparametric estimates of the hazard rate. It is feasible to do so because of the large sample size of our data (about 48,000). Figures 17 and 18 show such plots for R and V, respectively.

The procedure we use to calculate and plot the figures is as follows. For each interval of length δ, the estimate of the hazard rate is calculated as

$$\frac{[\,\#\text{ of events during } (t, t+\delta]\,]}{[\,\#\text{ at risk at } t]\ \times\ \delta}\ .$$

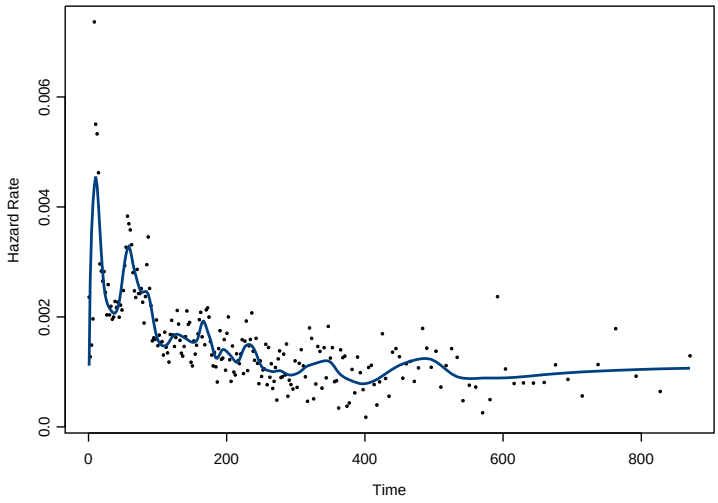
For smaller time values, t, the numbers at risk and event rates are large, and we let $\delta = 1$ second. For larger times, when fewer are at risk, larger δ's are used. Specifically, the larger intervals are constructed to have an estimated expected number of events per interval of at least four. Finally,

the hazard rate for each interval is plotted at the interval's median.

The curves superimposed on the plotted points are fitted using nonparametric regression. In practice we used LOCFIT (Loader, 1999), though other techniques, such as kernel procedures or smoothing splines, would yield similar fits. In particular, we have experimented with HEFT (Kooperberg, Stone and Truong, 1995), as can be seen in Mandelbaum et al. (2000). The smoothing bandwidth was chosen manually to provide a reasonably smooth estimate which, nevertheless, provided a satisfying visual fit to the data. We experimented with fitting techniques that varied the bandwidth to take into account the increased variance and decreased density of the estimates with increasing time. However, with our data these techniques had little effect and so are not used here.

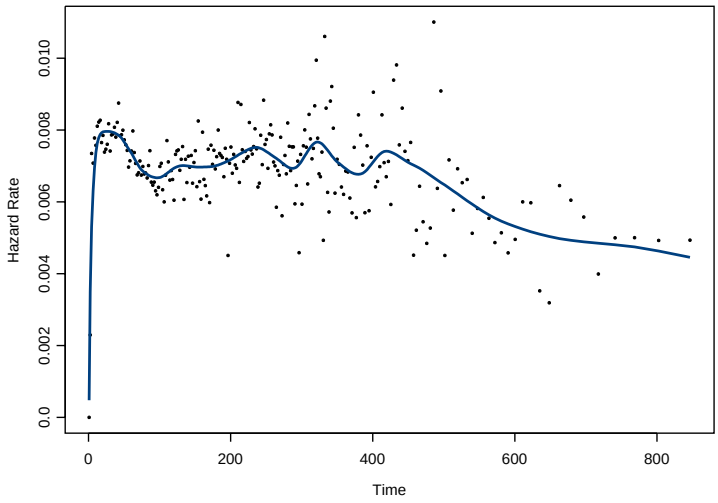
Figure 17 plots the hazard rates of the time willing to wait for regular (PS) calls. Note that it shows two main peaks. The first occurs after only a few seconds. When customers enter the queue, a "Please wait" message, as described in Section 3, is played for the first time. At this point some customers who do not wish to wait probably realize they are in queue and hang up. The second peak occurs at about $t = 60$, about the time the system plays the message again. Apparently, the message increases customers' likelihood of hanging up for a brief time thereafter, an effect that may be contrary to the message's intended purpose (or, maybe not!).

Figure 17: Hazard rate for the time willing to wait for PS calls (Nov.–Dec.)



In Figure 18, the hazard rate for the virtual waiting times is estimated for *all* calls (the picture for each type is very similar). The overall plot reveals rather constant behavior beyond 100 seconds, which suggests exponentiality of the tail.

Figure 18: Hazard rate for virtual waiting time (Nov.–Dec.)



6.4 Patience Index

Customer (im)patience on the telephone is important, yet it has not been extensively studied. In the search for a better understanding of (im)patience, we have found a *relative* definition to be of use here.

Let the means of V and R be m_V and m_R, respectively. Then we define the *patience index* to be the ratio of m_V/m_R, the mean of the time a customer is willing to wait to the mean of the time he or she is needed to wait. With this definition, we find that the overall patience index is 5.65, and those for the various types are as follows: regular customers (PS) = 5.44, stock trading (NE) = 5.6, new customers (NW) = 2.36 and internet technical support (IN) = 2.01. Thus, stock trading customers are found to be the most patient, and regular customers are next. This implies that not only that NE customers are willing to wait longer, they are also more patient than PS customers.

While this patience index makes sense intuitively, its calculation requires the application of survival analysis techniques to call-by-call data. Such data may not be available in certain circumstances. Therefore, we wish to find an *empirical index* which will work as an auxiliary measure for the patience index.

For the sake of discussion, we assume that V and R are independent and exponentially distributed. As a consequence of these assumptions, one can demonstrate that

$$\text{Patience Index} \stackrel{\triangle}{=} \frac{m_R}{m_V} = \frac{\mathrm{P}(V < R)}{\mathrm{P}(R < V)} \ .$$

Furthermore, $\mathrm{P}(V < R)/\mathrm{P}(R < V)$ can be estimated by (# served)/(# abandoned), and we define

$$\text{Empirical Index} \stackrel{\triangle}{=} \frac{\text{\# served}}{\text{\# abandoned}}.$$

Both the numbers of served and of abandoned calls are very easy to obtain from either call-by-call data or more aggregated call-center management reports. We have thus derived an easy-to-calculate empirical measure from a probabilistic perspective.

Note that we can also derive the same measure from a statistical perspective. Under the censoring setup, the MLEs for m_V and m_R are respectively (time on test)/(# served) and (time on test)/(# abandoned), denoted as $\hat{m}_V$ and $\hat{m}_R$. Thus the usual plug-in MLE for patience index is

$$\frac{\hat{m}_R}{\hat{m}_V} = \frac{\text{\# served}}{\text{\# abandoned}},$$

which is the same as the empirical index defined above.

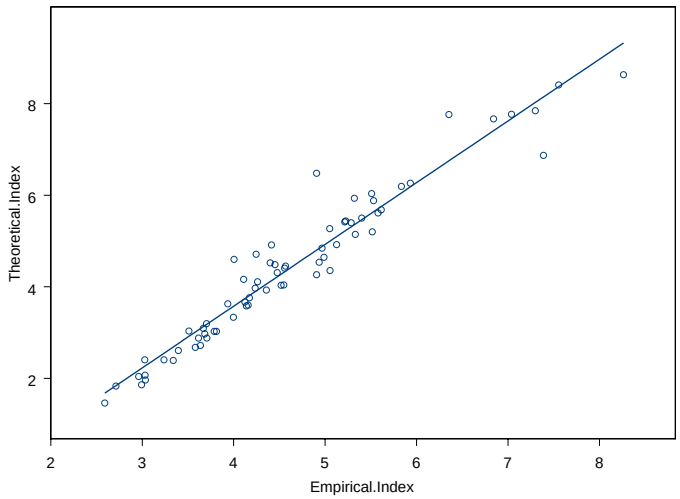
We can use our data to validate the empirical index as an estimate of the theoretical patience index. Recall, however, that the Kaplan-Meier estimate of the mean is biased when the last observation is censored or in the presence of heavy censoring. Nevertheless, a well known property of exponential distributions is that their quantiles are just the mean multiplied by certain constants, and we use quantiles when calculating the patience index. In fact because of heavy censoring, we sometimes do not obtain an estimate for the median or higher quantiles. Therefore, we used 1st quartiles when calculating the theoretical patience index.

Figure 19 shows how well the empirical index estimates the theoretical patience index. For each of 68 quarter hours between 7am and midnight, we calculated the 1st quartiles of V and R and their ratio, as well as the ratio of (# of served) to (# of abandoned). The figure plots the resulting 68 sample points, together with a least-squares fitted line superimposed. The linear relationship is almost perfect, with an R-square of 0.94. This result suggests that we can use the empirical measure as an index for human patience.

Recall that the linear relationship between the two indices is established under the assumption that R and V are exponentially distributed and independent. Here, however, the distribution for R is clearly not exponential, as shown in Figure 18. In addition, as discussed in Section 6.5, R and V may not be completely independent. Thus, we find the linear relation shown in Figure 19 to be astonishing.

Finally, we note another peculiar observation: the line does not have an intercept at 0 or a slope of 1, as suggested by the above theory. Rather, the estimated intercept and slope are -1.82 and 1.35, and are statistically different from 0 and 1. These differences may be due to the aggregation of data over different days at a fixed quarter hour. We are working on providing a theoretical explanation that accounts for these peculiar facts, as well as the fact that the two assumptions of exponentiality and independence do not hold for our data.

Figure 19: Patience index: empirical vs. theoretical



6.5 Dependence, or the violation of classical assumptions of survival analysis

A basic assumption in survival analysis is that failure times and censoring times are independent (or at least non-informative) of each other. Otherwise, estimation of the failure-time distribution could become biased.

In our data, such independence assumptions may be violated in many ways, however. As a start, the time that customers spend waiting is influenced by the abandonment behavior of those that precede them. Furthermore, a message (described above) informs customers about their positions in queue, possibly affecting their patience. Another source of dependence is repeated calling by the same customers: sometimes a customer may call many times over a short period of time. Finally, if the queue is long at a given time, then the virtual waiting time is likely to be long as well, and both are likely to be long also soon thereafter. If follows that virtual waiting times are dependent across customers and that, due to the message, there is dependence between R and V as well.

Having noted these violations of the standard assumptions of survival analysis, we nevertheless believe that our procedures are approximately valid. The robustness of the results, such as the fit shown in Figure 19, suggest that our approach is able to provide useful insights.

7 PREDICTION OF THE LOAD

This section reflects the view of the operations manager of a call center who plans and controls daily and hourly staffing levels. Prediction of the system "load" is a key ingredient in this planning. Statistically, this prediction is based on a combination of the observed arrival times to the system (as analyzed in Section 4), and service times during previous, comparable periods (as analyzed in Section 5).

In the following discussion we describe a convenient model and a corresponding method of analysis that can be used to give prediction confidence bounds for the load of the system. More specifically, we present a model in Section 7.4 for predicting the arrival rate and in Section 7.6 for predicting mean service time. In Section 7.7 we combine the two predictions to give a prediction (with confidence bounds) for the load according to the method discussed in Section 7.3.

7.1 Definition of load

In Section 4 we showed that arrivals follow an inhomogeneous Poisson process. We let $\Lambda_j(t)$ denote the true arrival rate of this process at time t on a day indexed by the subscript j. Figure 3 presents a summary estimate of $\bar{\Lambda}_\cdot(t)$, the average of $\Lambda_j(t)$ over weekdays in November and December.

For simplicity of presentation we treat together here all calls except the Internet calls (IN), since these were served in a separate system from August to December. The arrival patterns for the other types of calls appear to be reasonably stable from August through December. Therefore, in this section we use August–December data to fit the arrival parameters. To avoid having to adjust for the short-service-time phenomenon noted in Section 5.1, we use only November and December data to fit parameters for service times. Also, we consider here only regular weekdays (Sunday through Thursday) that were not full or partial holidays.

Together, an arbitrary arrival rate $\Lambda(t)$ and mean service time $\nu(t)$ at t define the "load" at that time, $L(t) = \Lambda(t)\nu(t)$. This is the expected time units of work arriving per unit of time, a primitive quantity in building classical queueing models, such as those discussed in Section 8.

Briefly, suppose one adopts the simplest Erlang-C (M/M/N) queueing model. Then if the load is a constant, L, over a sufficiently long period of time, the call center must be staffed, according to the model, with at least L agents; otherwise the model predicts that the backlog of calls waiting to be served will explode in an ever-increasing queue. Typically, a manager will need to staff the center at a staffing level that is some function of L – for example $L + c\sqrt{L}$ for some constant c – in order to maintain satisfactory performance. See Borst, Mandelbaum and Reiman (2000) and Garnett et al. (2002).

Determination of this function is a key goal of queueing theory, and its form depends on various

features of the call system and its desired performance levels. (See Section 8.) However, the first practical need is for an accurate prediction of L.

7.2 Independence of $\Lambda(t)$ and $\nu(t)$

In Section 5.4.4 we noted a qualitative similarity in the bi-modal patterns of arrival rates and mean service times shown in Figures 3 and 11. Three hypotheses that might lead to such a similarity are as follows:

Hypothesis A: The heavier-volume periods involve a different mix of customers, and that mix includes a higher proportion of customers that require lengthier service.

Hypothesis B: When the call volume is heavier, agents tend to handle calls more slowly. This may seem counterintuitive, but managers we have talked with suspect that longer service times may reflect a defensive reaction on the part of the agents. That is, during busy periods agents may respond to the pressure of a high volume of calls by slowing down.

Hypothesis C: Naturally, the heavier-volume periods are accompanied by a higher percentage of customers that abandon from the queue. It may be that this abandonment occurs predominantly among customers that only require relatively short service times. (If your business isn't very important you may not be willing to wait so long in queue.) Among the serviced calls, this would leave a mix containing a higher proportion of customers requiring longer services.

Under the first hypothesis it would be very reasonable to assume that $\Lambda(t)$ and $\nu(t)$ are conditionally independent, given the time of day. Such an assumption would be less reasonable under either of the other two hypotheses, although even then one would not expect $\Lambda(t)$ and $\nu(t)$ to be heavily dependent conditionally given the time of day.

Under the first hypothesis, it is also true that the load at any time would be independent of the number of agents currently operating. This is the prevailing assumption within standard queueing models. Under the latter two hypotheses, the load might be slightly sensitive to the number of agents working at that time. If there were a noticeable sensitivity, this would require modification of standard queueing models used for highly utilized systems.

We calculated the sample correlation, R, between the number of calls arriving in a 15 minute interval and the average service time of the calls served during that 15 minute interval, conditional on the time of day. For this calculation we used the data from November and December for all types of calls except Internet calls. IN calls were excluded since they have such a different arrival pattern and service-time distribution. The resulting value of R^2 was quite small (~ 0.047) but statistically significant in view of the large amount of data (3,391 quarter-hour intervals corresponding to 68 intervals on each of 44 days, with one interval excluded because of a malfunctioning system). We

also calculated the sample R^2 value when the mean of the log (service time) over the quarter hour intervals was used, rather than the mean of the service time itself. (The justification for doing this is that the service time is lognormal.) This R^2 value was only slightly larger (~ 0.060), and again it was highly statistically significant.

In both cases the sample value of R was negative. ($R = -0.22$ and -0.25 respectively.) This suggests that Hypothesis A is the correct explanation among the three hypotheses, above, since Hypotheses B and C imply positive values of R. Furthermore, it appears that, at any given time of day, the service agents may tend to work a little faster when the system is more congested. This is the opposite of Hypothesis B. However, this effect is quite weak, if it is present at all. Therefore, it seems reasonable to apply standard theory that assumes service times are conditionally independent of the load, given the time of day. It also seems reasonable to proceed as if $\Lambda(t)$ and $\nu(t)$ are independent processes, and we do so in the following calculations.

7.3 Coefficient of Variation for the prediction of $L(t)$

We discuss below the derivation of approximate confidence intervals for $\Lambda(t)$ and $\nu(t)$ that are based on observations of quarter-hour groupings of the data. The load, $L(t)$, is a product of these two quantities. Hence, exact confidence bounds are not readily available from individual bounds for each of $\Lambda(t)$ and $\nu(t)$. As an additional complication, the distributions of the individual estimates of these quantities are not normally distributed. Nevertheless one can derive reasonable approximate confidence bounds from the coefficient of variation (CV) for the estimate of L.

As a matter of general terminology, if W is any non-negative random variable with finite, positive, mean and variance, then its CV is defined by

$$CV(W) = \frac{SD(W)}{E(W)}.$$

If U and V are two independent variables and $W = UV$ then an elementary calculation yields

$$CV(W) = \sqrt{CV^2(U) + CV^2(V) + CV^2(U) \cdot CV^2(V)}.$$

In our case U and V correspond to Λ and ν. Predictions for Λ and ν are discussed in Sections 7.4 and 7.6. As noted above these predictions can be assumed to be statistically independent. Also, their CVs are quite small (under 0.1). Note that $\hat{L}(t) = \hat{\Lambda}(t)\hat{\nu}(t)$ and using standard asymptotic normal theory we can approximate $CV(\hat{L})(t)$ as

$$CV(\hat{L})(t) \approx \sqrt{CV^2(\hat{\Lambda})(t) + CV^2(\hat{\nu})(t)}.$$

This leads to approximate 95% confidence intervals of the form $\hat{L}(t) \pm 2\hat{L}(t)CV(\hat{L})(t)$. The constant 2 is based on a standard asymptotic normal approximation as being approximately 1.96.

7.4 Prediction of $\Lambda(t)$

Section 4 investigated the possibility of modeling the parameter Λ as a deterministic function of time of day, day of week and type of customer, and it rejected such a model. Here we construct a random-effects model that can be used to predict Λ and to construct confidence bands for that prediction. The model that we construct includes an autoregressive feature that incorporates the previous day's volume into the prediction of today's rate.

The arrival patterns appear to be relatively stable from August through December, and so our predictions are constructed on the basis of the August–December weekday data. In the model, which will be elaborated on below, we predict the arrival on a future day using arrival data for all days up to this day. Such predictions should be valid for future weekdays on which the arrival behavior follows the same pattern as those for this period of data. (Use of August–December data, rather than just the November–December data used in Figure 3 and 4, allows for more precise estimates of the arrival patterns.)

Our method of accounting for dependence on time and day is more conveniently implemented with balanced data, although it can also be used with unbalanced data. For convenience we have thus used data from only regular (non-holiday) weekdays in August through December on which there were no quarter-hour periods missing and no obvious gross outliers in observed quarter-hourly arrival rates. This leaves 101 days. For each day (indexed by $j = 1, \ldots, 101$) the number of arrivals in each quarter hour from 7am through 12 midnight was recorded as N_{jk}, $k = 1, \ldots, 68$. As noted in Section 4, these are assumed to be Poisson with parameter $\Lambda = \Lambda_{jk}$.

One could build a fundamental model for the values of Λ according to a model of the form

$$N_{jk} = \text{Poiss}(\Lambda_{jk}), \;\; \Lambda_{jk} = R_j \tau_k + \varepsilon'_{jk}, \tag{5}$$

where the τ_k are fixed deterministic quarter-hourly effects, the R_j denote random daily effects with a suitable stochastic character, and the ε'_{jk} are (hopefully small) random errors. Note that this multiplicative structure is natural, in that the τ_k's play the role of the expected proportion of the day's calls that fall in the k-th interval. This is assumed to not depend on the R_j's, the expected overall number of calls per day. (We accordingly impose the side condition that $\sum \tau_k = 1$.)

We will, instead, proceed in a slightly different fashion that is nearly equivalent to (5), but is computationally more convenient and leads to a conceptually more familiar structure. The basis for our method is a version of the usual variance stabilizing transformation. If X is a Poiss(λ) variable then $V = \sqrt{X + \frac{1}{4}}$ has approximately a mean $\theta = \sqrt{\lambda}$ and variance $\sigma^2 = \frac{1}{4}$. This is nearly precise even for rather small values of λ. (One could instead use the simpler form $\sqrt{X}$ or the version of Anscombe (1948) that has $\sqrt{X + \frac{3}{8}}$, in place of $\sqrt{X + \frac{1}{4}}$; only numerically small changes would result. Our choice is based on considerations in Brown et al. (2001).) Additionally, V is asymptotically normal (as $\lambda \to \infty$), and it makes sense to treat it as such in the models that

follow. We thus let $V_{jk} = \sqrt{N_{jk} + \frac{1}{4}}$, and assume the model

$$\begin{aligned} V_{jk} &= \theta_{jk} + \varepsilon^*_{jk} \quad \text{with} \quad \varepsilon^*_{jk} \overset{iid}{\sim} N\left(0, \tfrac{1}{4}\right), \\ \theta_{jk} &= \alpha_j \beta_k + \varepsilon_{jk}, \\ \alpha_j &= \mu + \gamma V_{j-1,+} + A_j, \end{aligned} \tag{6}$$

where $A_j \sim N(0, \sigma_A^2)$, $\varepsilon_{jk} \sim N(0, \sigma_\varepsilon^2)$, $V_{j,+} = \sum_k V_{jk}$, and A_j and ε_{jk} are independent of each other and of values of $V_{j',k}$ for $j' < j$. Note that α_j is a random effect in this model. Furthermore the model supposes a type of first order auto-regressive structure on the random daily effects. Models with an auto-regressive component similar to this are common in call center contexts. The correspondence between (5) and (6) implies that this structure corresponds to an approximate assumption that

$$R_j = \left(\gamma \sum_k \sqrt{N_{j-1,k} + \frac{1}{4}} + A_j \right)^2.$$

The model is thus not quite a natural one in terms of the R_j, but it appears more natural in terms of the V_{jk} in (6) and is computationally convenient.

The parameters γ and β_k need to be estimated, as well as μ, σ_A^2 and σ_ε^2. We impose the side condition $\sum \beta_k^2 = 1$, which corresponds to the condition $\sum \tau_k = 1$. The goal is then to derive confidence bounds for $\theta_{jk} = \sqrt{\Lambda_{jk}}$ in (6), and by squaring the bounds we obtain corresponding bounds for Λ_{jk}.

The parameters in the model (6) can easily be estimated by a combination of least-squares and method-of-moments. Begin by treating the $\{\alpha_j\}$'s as if they were fixed effects and using least-squares to fit the model

$$V_{jk} = \alpha_j \beta_k + \left(\varepsilon_{jk} + \varepsilon^*_{jk}\right).$$

This is an easily solved nonlinear least squares problem. It yields estimates $\hat{\alpha}_j$, $\hat{\beta}_k$ and $\hat{\sigma}^2$, where the latter estimate is the mean square error from this fit. σ_ε^2 can then be estimated by method-of-moments as

$$\hat{\sigma}_\varepsilon^2 = \hat{\sigma}^2 - \frac{1}{4}.$$

Then use the estimates $\{\hat{\alpha}_j\}$ to construct estimates of the parameters μ, γ and σ_A^2 that appear in the auto-regressive part of the model. Thus, use the "observations" $\{\hat{\alpha}_j\}$ to construct the least squares estimates of these parameters that would be appropriate for a linear model of the form

$$\hat{\alpha}_j = \mu + \gamma V_{j-1,+} + \varepsilon^{**}_j. \tag{7}$$

This yields least-squares estimates, $\hat{\mu}$ and $\hat{\gamma}$, and the standard mean square error estimator $\hat{\sigma}^{**2}$ for the variance of ε^{**}_j.

The estimates calculated from our data for the quantities related to the random effects are

$$\begin{aligned} &\hat{\mu} = 97.88,\ \ \hat{\gamma} = 0.6784 \ \text{(with corresponding } R^2 = 0.501), \\ &\hat{\sigma}^{**2} = 408.3,\ \ \hat{\sigma}^2_\varepsilon = 0.1078 \ \text{(since } \hat{\sigma}^2 = 0.3578). \end{aligned} \tag{8}$$

The value of R^2 reported here is derived from the estimation of γ in (7). This value thus measures the reduction in sum of squared error due to fitting the $\{\hat{\alpha}_j\}$ by this model involving previous day's call volumes, $V_{j-1,+}$. The large value of R^2 in (8) makes it clear that the introduction of the auto-regressive model (7) noticeably reduces the prediction error (by about 50%) relative to that obtainable from a model with no such component (i.e., one in which a model of the form (6) holds with $\gamma = 0$).

For a prediction, $\breve{\Lambda}_k$, of tomorrow's value of Λ_k at a particular quarter hour (indexed by k), one would use the above estimates along with today's value of V_+. From (6) it follows that tomorrow's prediction is

$$\breve{\theta}_k = \hat{\beta}_k \, (\hat{\gamma} V_+ + \hat{\mu}) \tag{9}$$

as an estimate of

$$\theta_k = \beta_k \, (\gamma V_+ + \mu + \varepsilon^{**}) + \varepsilon \tag{10}$$

where $\varepsilon^{**} \sim N(0,\, \sigma^{**2})$ and $\varepsilon \sim N(0,\, \sigma^2_\varepsilon)$ are independent. The variance of the term in parentheses in (9) is the prediction variance of the regression in (7). Denote this by $\text{PredVar}(V_+)$. The coefficient of variation of $\hat{\beta}_k$ turns out to be numerically negligible compared to other coefficients of variation involved in (9) and (10). Hence

$$\text{Var}(\breve{\theta}_k) \approx \hat{\beta}^2_k \times \text{PredVar}(V_+) + \hat{\sigma}^2_\varepsilon. \tag{11}$$

These variances can be used to yield confidence intervals for the predictions of θ_k. The bounds of these confidence intervals can then be squared to yield confidence bounds for the prediction of Λ_k. Alternatively one may use the convenient formula $CV(\breve{\theta}^2_k) \approx 2 \times CV(\breve{\theta}_k)$, and produce the corresponding confidence intervals.

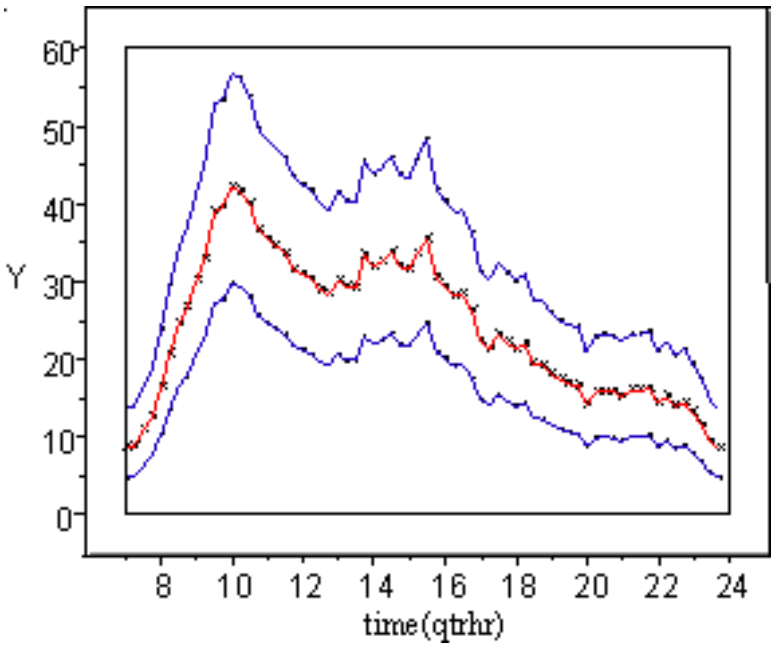
The plot in Figure 20 shows the confidence bounds for the predictions of Λ_k on a day following one having $V_+ = 340$. (This is a heavy volume day, since our data has $\overline{V}_+ = 306$. Calculation yields $\text{PredVar}(340) = 417.8$.)

Note that the values of $CV(\breve{\theta}^2_k)$ here are in the range of 0.25 (for early morning and late evening) down to 0.16 (for mid-day). Note also that both parts of (11) are important in determining variability – the values of $\text{Var}(\breve{\theta}_k)$ range from 0.14 (for early morning and late evening) up to 0.27 (for mid-day). The fixed part of this is $\hat{\sigma}^2_\varepsilon = 0.11$, and the remainder results from the first part of (11), which reflects the variability in the estimate of the daily volume figure, A^*, in (7).

Correspondingly, better estimates of daily volume (perhaps based on covariates outside our data set) would considerably decrease the CVs during mid-day but would not have much effect on those

for early morning and late evening. (Incidentally, we tried including day of the week (Sunday, Monday, etc.) as an additional covariate in the model (6), but with the present data this did not noticeably improve the resulting CVs.)

Figure 20: 95% Prediction intervals for Λ derived via (10)-(11) when $V_+ = 340$



7.5 Diagnostics for the model for $\Lambda(t)$

The model in Section 7.4 is built from several assumptions of normality. These can be empirically checked in the usual way by examining residual plots and Q-Q plots of residuals. All the relevant diagnostic checks showed good fit to the model. For example, the Q-Q plots related to A^* and ε^{**} support the normality assumptions in the model.

According to the model the residuals corresponding to ε_{jk} should also be normally distributed. The Q-Q plot for these residuals has slightly heavier-than-normal tails; but only 5 (out of 6,868) values seem to be heavily extreme. These heavy extremes correspond to quarter-hour periods on different days that are noticeably extreme in terms of their total number of arrivals.

7.6 Prediction of $\nu(t)$

In order to combine in Section 7.7 the estimates of $\nu(t)$ derived here with the estimates of $\Lambda(t)$ derived in Section 7.4, we also model the service time according to quarter hour intervals in this

section. In other respects the model developed in this section resembles the nonparametric model of Section 5.4.

To avoid having to model the short-service-time phenomenon discussed in Section 5.1, we use weekday data from only November and December. The lognormality discussed in Section 5.3 allows us to model log(service times), rather than service times. Therefore, we let Y_{jkl} denote the log(service time) of the l-th call served by an agent on day j, $j = 1, \ldots, 44$, in quarter-hour interval k, $k = 1, \ldots, 68$. In total there are $n = 57,152$ such calls. (We deleted the few call records showing service times of 0 or of > 3600 seconds.)

For purposes of prediction we will ultimately adopt a model like (3) of Section 5.4, namely

$$Y_{jkl} = \mu + \kappa_k + \varepsilon_{jkl}, \quad \varepsilon_{jkl} \sim N(0, \sigma_k^2) \text{ (indep.)}. \tag{12}$$

However, before adopting such a model we investigate whether there are day-to-day inhomogeneities that might yield a better prediction model. (Such is the case in the model (6), above, for arrival rates, where the terms involving A_j capture this inhomogeneity.) Hence we investigate a mixed model of the form

$$Y_{jkl} = \mu + B_j + \kappa_k + \varepsilon_{jkl}. \tag{13}$$

Note that in this case an additive model is standard and statistically natural for such a Gaussian nonparametric regression setting, rather than a multiplicative one such as (5). It is also convenient to analyze. Here, $\mu + B_j$ represents the daily mean service time for day j.

A least-squares analysis of the model (13) yields a partial $R^2 = 0.005$ for the terms involving B_j. This is statistically significant (P-value < 0.0001). (Remember, $n = 57,152$ is very large!) But it has very little numerical importance: even **if** the B_j could be accurately predicted, this knowledge would reduce the mean square prediction error only by a very small factor. In fact, the B_j cannot be predicted with complete accuracy. We conclude there is no practical advantage in basing predictions on a model of the form (13) rather than (12), and we use (12) in what follows. (We also investigated a model that used the day – Sunday, Monday, etc. – as an additional factor but found no useful information in doing so.)

The goal is to produce a set of confidence intervals (or corresponding CVs) for the parameter

$$\nu_k = \exp\left(\mu + \kappa_k + \frac{\sigma_k^2}{2}\right). \tag{14}$$

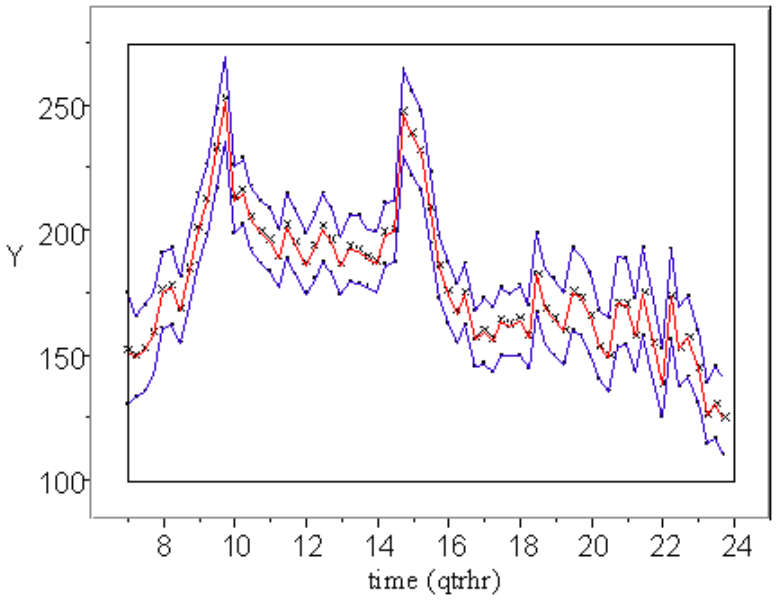
The basis for this is contained in Section 5.4, except that here we use estimates from within each quarter-hour time period, rather than kernel smoothed estimates. The most noticeable difference is that the standard error of σ_k^2 is now estimated by

$$\mathrm{se}_{\sigma_k^2} \approx \sqrt{\frac{2}{n_k - 1}} S_k^2, \tag{15}$$

where n_k denotes the number of observations within the quarter hour indexed by k and S_k^2 denotes the corresponding sample variance from the data within this quarter hour. This estimate is motivated by the fact that if $X \sim \mathrm{N}(\mu, \sigma^2)$ then $\mathrm{Var}((X-\mu)^2) = 2\sigma^4$.

Figure 21 is the resulting plot of predictions of ν_k and their confidence intervals. It is qualitatively very similar to that of Figure 11, except more jagged, because it is not based on nonparametric regression smoothing. It is also somewhat different because we have included all but the IN calls along with the regular (PS) calls.

Figure 21: Estimates and 95% confidence intervals for ν_k



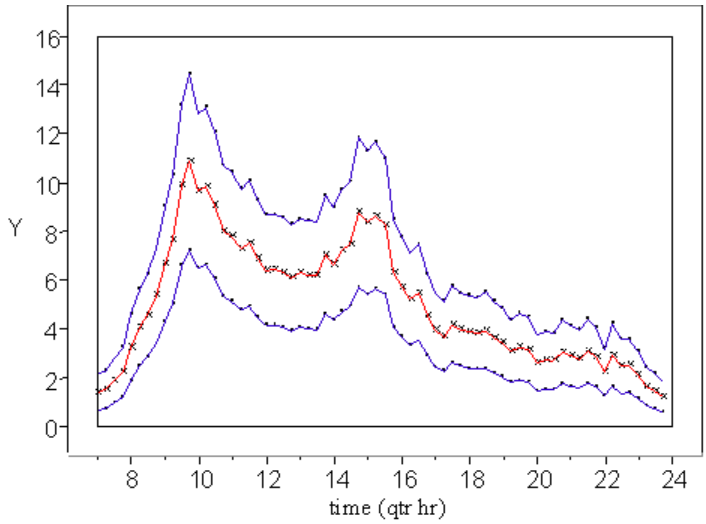
Coefficients of Variation for these estimates can be calculated from the approximate (Taylor series) formula $CV^*(\hat{\nu}_k) \approx CV\left(\hat{\mu} + \hat{\kappa}_k + \frac{\hat{\sigma}_k^2}{2}\right)$. (The intervals $\hat{\nu}_k \pm 1.96 \times \hat{\nu}_k \times CV^*$ agree with the above to within 1 part in 200, or better.) The values of CV here range from 0.03 to 0.08. This is much smaller than the corresponding values of CVs for estimating $\Lambda(t)$. Consequently in producing confidence intervals for the load, $L(t)$, the dominant uncertainty is that involving estimation of $\Lambda(t)$.

7.7 Confidence intervals for $L(t)$

The confidence intervals in Figures 20 and 21 can be combined as described in Section 7.3 to obtain confidence intervals for L in each quarter hour period. Care must be taken to first convert the estimates of Λ and ν to suitable, matching units. Figure 22 shows the resulting plot of predicted load on a day following one in which the arrival volume had $V_+ = 340$.

The intervals in Figure 22 are still quite wide. This reflects the difficulty in predicting the load at

Figure 22: 95% prediction intervals for the load, L, following a day with $V_+ = 340$.



a relatively small center such as ours. We might expect that predictions from a large call-center would have much smaller CVs, and we are currently examining data from such a large center to see whether this is the case. Of course, inclusion (in the data and corresponding analysis) of additional informative covariates for the arrivals might improve the CV's in a plot like Figure 22.

8 SOME APPLICATIONS OF QUEUEING SCIENCE

In Section 2, we noted a distinction between queueing *theory* and queueing *science*. Queueing theory concerns the development of formal, mathematical models of congestion in stochastic systems, such as telephone and computer networks. It is a highly-developed discipline that has roots in the work of A. K. Erlang (Erlang 1911, 1917) at the beginning of the 20th century. Queueing science, as we view it, is the theory's empirical complement: it seeks to validate and calibrate queueing-theoretic models via data-based scientific analysis. In contrast to queueing theory, however, queueing science is not so highly developed. While there exist scattered applications in which the assumptions of underlying queueing models have been checked, we are not aware of a systematic effort to validate queueing-theoretic results.

One area in which extensive work *has* been done – and has motivated the development of new theory – involves the arrival times of internet messages (or message packets). See for example, Willinger, Taqqu, Leland and Wilson (1995), Cappe, Moulines, Pesquet, Petropulu and Yang (2002) and the references therein. These arrivals have been found to involve heavy tailed distributions and/or long

range dependencies (and thus differ qualitatively from the results reported in our Section 4).

Queueing-scientific analyses are of use in understanding whether an existing model is adequate for application in a particular setting and, if not, how it should be augmented to better capture system behavior. Conversely, empirical results sometime highlight unanticipated usefulness of models that are known not to capture certain system features, and they stimulate the search for a theoretical explanation for the models' robustness.

In this section, we use our call-center data to produce two examples of this type of analysis. In §8.1 we validate (and refute) some classical theoretical results. In §8.2, we demonstrate the robustness (and usefulness) of a relatively *simple* theoretical model, namely the M/M/N+M (Erlang-A) model, for performance analysis of a *complicated* reality, namely our call center.

8.1 Validating Classical Queueing Theory

We analyze two congestion laws: first, the relationship between patience and waiting, which is a byproduct of the Little's law (Zohar et al., 2002); then, the interdependence between service quality and efficiency, as it is manifested through the classical Khintchine-Pollaczek formula (see, for example, Equation (5.68) in Hall 1991).

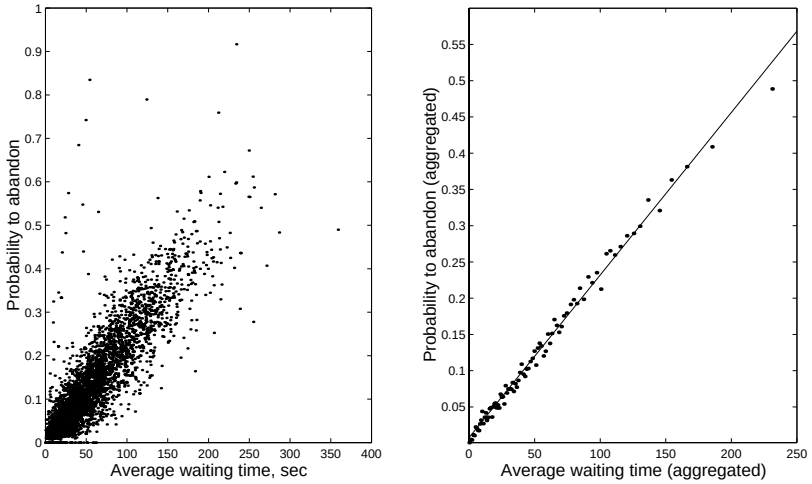
On Patience and Waiting: The probability of abandonment or, in practice, the fraction of customers who abandon, is perhaps the most important operational measure for call center performance. Clearly, the shorter the delay in queue, the fewer people who abandon due to impatience. But similar statements can be made about most performance measures: in general less wait implies better service.

What distinguishes abandonment from the rest is the fact that it is one of only a few measures through which customers indirectly "inform" the call center on their *subjective* view of whether its service is "worth the wait." (Other such measures are various retrial rates, for example after abandonment or after being served). Commonly used measures, such as average waiting times, are also important, since they relate to customers' service experiences, but they are *objective* at the system level. It is thus theoretically important and practically useful that there exists a simple functional relationship between the probability of abandonment and the average wait. This will be now demonstrated.

Figure 23 depicts the relationship between average waiting time (objective measure) and the probability of abandonment (subjective measure). It was plotted using the yearly data of 1999. First, the probability of abandonment and average wait were computed for the 3,867 hourly intervals that constitute the year. The left plot of Figure 23 presents the resulting "cloud" of points as they scatter on the plane. For the right plot we use a procedure that is designed to emphasize dominating patterns.

This procedure, which is used in later plots as well, is as follows. The hour intervals are ordered according to their average waiting times, and adjacent groups of 45 points are aggregated (further averaged): this forms the 86 points produced in the right plot of Figure 23. (The last point of the aggregated plot is an average of 42 hour intervals.)

Figure 23: Probability of abandonment versus average waiting time



The linear fit that comes out of the second graph is remarkable. Indeed, if W is the waiting time and R is the time a customer is willing to wait (referred to as *patience* in the sequel), the law

$$\% \text{ Abandonment} = \frac{\mathrm{E}(W)}{\mathrm{E}(R)} \tag{16}$$

is provable for models with *exponential* patience (as in Baccelli and Hebuterne 1981 or Zohar et al. 2002 for example). But this obviously is *not* the case here. (See Figure 17 for the hazard rate of patience, which is far from being constant, as it should be if R is exponential.)

Thus, the need arises for a theoretical explanation of why this linear relationship holds in models with *generally distributed* patience. (Such an explanation is currently being pursued.) Similarly, the identification and analysis of situations in which non-linear relationships arise remains an interesting open question worthy of further research.

A rough estimate of average patience, under the hypothesis of its exponentiality and guided by equation (16), is the inverse of the regression-line slope, which is 446 seconds in our case. (The regression results for the data without aggregation turn out to be nearly the same.) This is somewhat smaller than the estimates provided in Table 5, namely 535–806 seconds, and the discrepancy can be, perhaps, attributed to the non-exponentiality of patience.

On Efficiency and Service Levels: As fewer agents cope with a given workload, operational efficiency increases. The latter is typically measured by the system (or agents') occupancy, namely

the average utilization of agents over time. Formally, it is defined as

$$\rho \;=\; \frac{\lambda}{N\mu}\,, \tag{17}$$

where λ is the arrival rate, $\mathrm{E(S)} = 1/\mu$ is the average service time, and N number of active agents, either serving customers or available to do so. Thus, the staffing level – the number of available agents N – is required to calculate agents' occupancy. Neither occupancies nor staffing levels are explicit in our database, however.

We have experimented with several algorithms for estimating staffing levels and agents' occupancy and have ultimately settled on the following recursive procedure. Define

$$\begin{aligned} N_t &= \text{number of agents working at time } t; \\ Q_t &= \text{number of calls in queue at time } t; \text{ and} \\ H_t &= \text{number of calls being served at time } t. \end{aligned} \tag{18}$$

Note that both Q_t and H_t are exactly calculable from our database. Evidently, $N_t \geq H_t$, and our goal is to estimate their difference. We then propose:

$$\begin{aligned} &\text{if } Q_t > 0, \text{ let } N_t = H_t; \text{ otherwise} \\ &\text{if } Q_t = 0, \text{ let } N_t = \max(N_{t-1}, H_t), \end{aligned} \tag{19}$$

where t advances in increments of one minute, and $N_{7{:}00am} = 0$. We note that the first part of the rule (19) reflects a "work-conservation" assumption that is common in queueing systems: if there is a queue, then every available agent is busy. The second part of the rule is based on $N_t \geq H_t$ and on our decision to estimate the staffing level at time t by the value at the previous time point $t-1$, unless information concerning its change has become available.

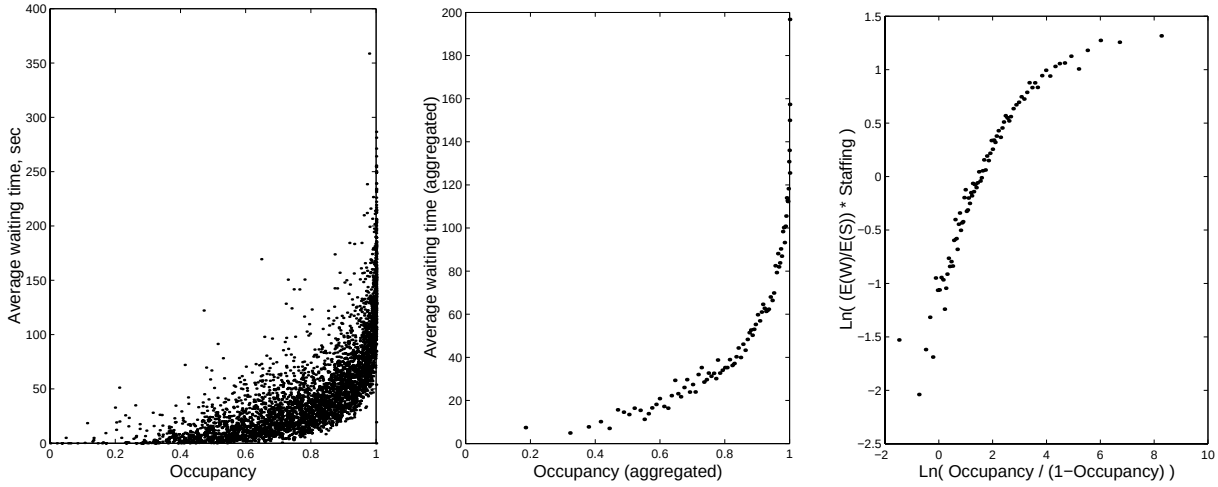
Agents' occupancy during a specific hour can be then computed by dividing the hourly average number of calls in service, equal to the average number of agents providing service, by the average staffing level. The latter includes both idle agents and those providing service. Observe that, over periods longer than one minute, the average number of agents providing service need not be integral.

The three plots of Figure 24 depict the relationship between average waiting time and agents' occupancy. The first plot shows the result for each of the 3,867 hourly intervals over the year. In the second, the method of data aggregation used in Figure 23 is repeated. (We aggregate neighboring points on the x-axis, namely, occupancy.)

The pattern of the first two plots is reminiscent of the classical Khintchine-Pollaczek formula, or more precisely its M/G/N approximations for the average time in queue. Here we use the simplest approximation (Whitt 1993):

$$\mathrm{E}(W) \approx \frac{1}{N} \cdot \frac{\rho}{1-\rho} \cdot \frac{1+c_s^2}{2} \cdot \mathrm{E}(S), \tag{20}$$

Figure 24: Agents' occupancy vs. average waiting time



where c_s is the coefficient of variation of the service time. Note that we calculate the agents' occupancy, ρ, as $\rho = \dfrac{\lambda_{eff}}{N\mu}$, where λ_{eff} is the *effective* arrival rate, namely the arrival rate of customers who get served (that is, those who do not abandon).

The formula is exact for the M/G/1 queue, and it always provides an upper bound on E(W) in the G/G/1 queue, a system with Generally distributed interarrival and service times. Furthermore, the upper bound becomes asymptotically tight in 'heavy traffic' as $\rho \uparrow 1$. See Iglehart and Whitt (1970) and Whitt (2002). For more refined approximations, see Sze (1984) and Whitt (1993).

The third plot of Figure 24 tests the applicability of the Khintchine-Pollaczek formula in our setting by plotting $N \cdot \mathrm{E}(W)/\mathrm{E}(S)$ versus $\rho/(1-\rho)$. To check if the two exhibit the linear pattern implied by (20), we display an aggregated version of the data as a scatter plot on a logarithmic scale. (The aggregation of hourly data is performed as before.) We believe that variability of the coefficient of variation c_s over hourly intervals does not affect the conclusions below. (In general, c_s turns out to be somewhat larger than 1 for various subsets of our data; see Table 2 for example.)

The graph pattern is *not* linear. This can be explained by the fact that classical versions of Khintchine-Pollaczek formula are not appropriate for queueing systems with abandonment. Indeed, for high levels of utilization (heavy traffic), abandonment strongly reduces average wait. Specifically, Khintchine-Pollaczek implies that $\mathrm{E}(W) \to \infty$, as $\rho \uparrow 1$. However, in models with abandonment, the average wait would increase to the average patience with $\rho \uparrow 1$. As for light traffic, Khintchine-Pollaczek is only a rough upper bound.

Note that queueing systems with abandonment usually give rise to dependence between successive interarrival times of *served* customers, as well as between interarrival times of served customers

and service times. For example, long service times could entail massive abandonment and, therefore, long interarrival times of served customers. A version of the Khintchine-Pollaczek formula that ccan potentially accommodate such dependence is derived in Fendick, Saksena and Whitt (1989). Theoretical research is needed to support the fit of these latter results to our setting with abandonment.

8.2 Fitting the M/M/N+M model (Erlang-A)

The Erlang-C (M/M/N) model is by far the most common theoretical tool used in the practice of call centers. In our judgment, its prevalence stems from its simplicity in that only three easily-estimable parameters are needed for its use: an arrival rate λ, a service rate μ and the number of accessible agents N. The parameters λ and μ are fit to, say, each hour of a day, and the model is then used to determine the least number of agents N that guarantees a given level of service.

Recall that the M/M/N model assumes Poisson arrivals at rate λ, exponentially distributed service times with mean $1/\mu$, and N agents. In the M/M/N model customers who are delayed in queue remain waiting until they are served; there is no abandonment. Increased focus on service levels has heightened industry awareness of and interest in customer abandonment, however, and it calls for the application of models which better capture abandonment effects.

The M/M/N+M model (Palm 1943) is the simplest abandonment-sensitive refinement of the Erlang-C (M/M/N) system. It is an Erlang-C model onto which exponentially distributed, or Markovian, customer patience (time to abandonment) is added, hence the '+M' notation. In addition to the Erlang-C's inputs (λ, μ, N), it requires an estimate of the average duration of customer patience, $1/\theta$, or equivalently an individual abandonment rate θ. Because it captures Abandonment behavior, we call M/M/N+M the "Erlang-A" model. See Garnett et al. (2002) for further details.

The analysis in Sections 5 and 6 shows that, in our call center, both service times and patience are *not* exponentially distributed, however. Nevertheless, it is our experience that simple models have often been found to be robust in describing complex systems. We therefore check whether the Erlang-A model provides a useful description of our *hourly* data.

Both positive and negative answers to this question would be valuable. If the model performs well, it can be used to help determine required staffing levels in our call center. Currently, it is impossible to find software that incorporates features such as lognormal service times, customer patience with the hazard-rate peaks, or multiple types of customers, and by including abandonment the Erlang-A model pushes forward the state of the art. If the fit is poor, we would reach the conclusion that available models are inadequate. In this case, improvement in the performance of commonly used Erlang-C-based staffing software may require alternatives to analytical models, such as simulation.

8.2.1 The Erlang-A Model

In testing the performance of the Erlang-A model, the philosophy that we adopt is reminiscent of that used in linear regression: given a set of useful models – in our case, Markovian models with homogeneous servers and a homogeneous population of impatient customers – one seeks model parameters that best fit the data. Here, we fit Erlang-A models by fine-tuning the patience of customers as they wait 'on hold' to be served.

More specifically, we use the following approach. First, we take arrival rates and average service times from specific subsets of the November data. We then ask: can we choose the "number of agents" and "average patience" that are input to an Erlang-A model so that output performance is close to that observed in our data? To calculate performance statistics we use *iProfiler* software (4CallCenters, 2002), which implements the Erlang-A model as follows.

Given constant arrival, service, and abandonment rates, as well a fixed, integral number of agents, the software uses an Erlang-A model to generate stationary measures of system performance, for example the fraction of customers delayed, the fraction of customers abandoning, average waiting time, agent occupancy and more. We emphasize that both the Erlang-A model and the *iProfiler* software require *whole* numbers of agents for each interval of time for which a model is fit.

Given the *iProfiler* results, we validate the fit by checking that the resulting average patience is close to the estimates calculated in Section 6. We similarly check that the number of agents (input) and agents' occupancy (output) are close to the estimates generated by the algorithm (19). Note, however, that for periods longer than one minute, the algorithm based on (19) can generate fractional numbers of agents. (For an interval of length T minutes that begins at t, $N = \frac{1}{T}\sum_{s=t}^{t+T-1} N_s$.) Thus, the "best fitted" N for *iProfiler* typically will not equal the N calculated via (19).

Erlang-A Analysis. Assessment of Individual Hours. We consider three time intervals: 11:00am–12:00pm, when there is a moderate workload; 12:00pm–1:00pm, when the workload is relatively light; and 3:00pm–4:00pm when there is a high workload. In each of the three analyses, we pool the data of the three service types – PS, NE and NW – that share the same queue. For each hour, the input to the *iProfiler* software consists of arrival rate, average service time, average patience, and number of agents. The output consists of various performance characteristics such as average waiting time, probability of wait, probability of abandonment and agent occupancy.

Table 6 summarizes our analysis. It is clear that the Erlang-A model provides a potentially very useful approximation for the performance characteristics of our call center. Indeed, the value of patience that comes out is 540 seconds, which is not far from the values given in Table 5. The correspondence between occupancy levels and agent numbers is also very good. In general, we observe a very good fit of most performance measures during 11:00am–12:00pm and a reasonable

fit for the other two periods.

Table 6: Validating the Erlang-A (M/M/N+M) model

Inputs to *iProfiler*	11am–12pm	12pm-1pm	3pm–4pm
Arrival Rate λ (per min.)	116	103	126
Avg. Service Time E(S) = $1/\mu$ (min.)	3:11	3:03	3:26
Avg. Patience $1/\theta$ (min.)	9	9	9
Number of Agents N	6	6	6
N Estimated from Data via (19)	5.92	5.76	5.93

Statistics from the data and from *iProfiler*	11am–12pm	12pm-1pm	3pm–4pm
Avg. Wait from Data (min.)	1:02.6	0:48.8	1:58.8
Avg. Wait from *iProfiler* (min.)	1:10.5	0:40.4	1:52.8
Fraction with Wait > 0 from Data	71.9%	58.3%	90.3%
P{Wait > 0} from *iProfiler*	68.5%	50.9%	84.0%
Fraction Abandoning from Data	13.9%	10.2%	23.4%
P{Abandon} from *iProfiler*	13.1%	7.5%	20.9%
Occupancy ρ Estimated from Data via (19)	90.6%	82.8%	95.9%
Occupancy ρ from *iProfiler*	89.2%	80.7%	95.1%

Note that most performance measures in Table 6 show a slight underestimation of the actual data values. We believe this underestimation is due, at least partially, to the fact that the number of agents N must be integer in *iProfiler*. Indeed, Table 6 shows that estimates for N arrived at via (19) vary between 5.76 and 5.93. Since the values of the performance measures in Table 6 increase as the number of agents decreases, most values in the "Data" rows are somewhat larger than the corresponding values in the "*iProfiler*" rows.

An improved fit can be obtained from *iProfiler* via simple linear interpolation. For concreteness, consider the time interval 12:00pm–1:00pm, where the estimate for average staffing level from (19) is 5.76. If one records two sets of *iProfiler* statistics, one for $N = 5$ and the other for $N = 6$, and interpolates, the results for some performance measures are as follows: average wait of 49.9 sec (vs. 40.4 sec for $N = 6$), probability of abandonment 9.3% (vs. 7.5% for $N = 6$) and probability of wait 56.0% (vs. 50.9% for $N = 6$). Thus, the interpolated statistics get much closer to the data values in Table 6.

It is also important to note that the M/M/N (Erlang-C) model which is typically used in practice is, in fact, completely inconsistent with our data. For example, if one ignores abandonment and uses the other values from Table 6, the Erlang-C model is useless for the periods 11:00am–12:00pm and 3:00pm–4:00pm. Indeed, at these times agent utilization turns out to be larger than 100%,

indicating that the queue "explodes" – the system can not reach stationarity.

Erlang-A Analysis. Overall Assessment. We now validate the Erlang-A model against the overall hourly data used in Section 8.1. Three performance measures are considered: probability of abandonment, average waiting time and probability of waiting (at all). Their values are calculated for our 3,867 hourly intervals using exact Erlang-A formulae (see details in the Remark below). Then the results are aggregated along the same method employed in Figures 23 and 24. The resulting 86 points are compared against the line $y = x$.

As before, the parameters λ and μ are easily computed for every hourly interval. For the overall assessment, we calculate each hour's average number of agents N via (19). Because the resulting Ns need not be integral, we apply a continuous extrapolation of the Erlang-A formulae, obtained from relationships developed in Palm (1943).

For θ, we use formula (16), valid for exponential patience, in order to compute 17 hourly estimates of $1/\theta = E(R)$ (for the 17 one-hour intervals 7am-8am, 8am-9am, ..., 11pm-12pm). The values for E(R) ranged from 5.1 min (8am-9am) to 8.6 min (11pm-12pm). We judged this to be better than estimating θ individually for each of the 3,867 hours (which would be very unreliable) or, at the other extreme, using a single value for all intervals (which would ignore possible variations in customers' patience over the time of day).

The results are displayed in Figure 25. The figure's two left-hand graphs exhibit a relatively small yet consistent overestimation with respect to empirical values, for moderately and highly loaded hours. (We plan to explore the reasons for this overestimation in future research.) The right-hand graph shows a very good fit everywhere, except for very lightly and very heavily loaded hours. The underestimation for small values of P{Wait} can be probably attributed to violations of work conservation (idle agents do not always answer a call immediately). Summarizing, it seems that these Erlang-A estimates can be used as reasonable *upper bounds* for the main performance characteristics of our call center.

8.2.2 Approximations

Garnett et al. (2002) develops approximations of various performance measures for the Erlang-A (M/M/N+M) model. Such approximations require significantly less computational effort than exact Erlang-A formulae. The theoretical validity of the approximation is established in Garnett et al. (2002) for large Erlang-A systems. While this is not exactly our case, Figure 26, below, nevertheless demonstrates a good fit between data averages and the approximations.

In fact, the fits for the probability of abandonment and average waiting time are superior to those in Figure 25 (the approximations provide somewhat larger values than the exact formulae). This phenomenon suggests two interrelated research questions of interest: explaining the overestima-

Figure 25: Erlang-A formulas vs. data averages

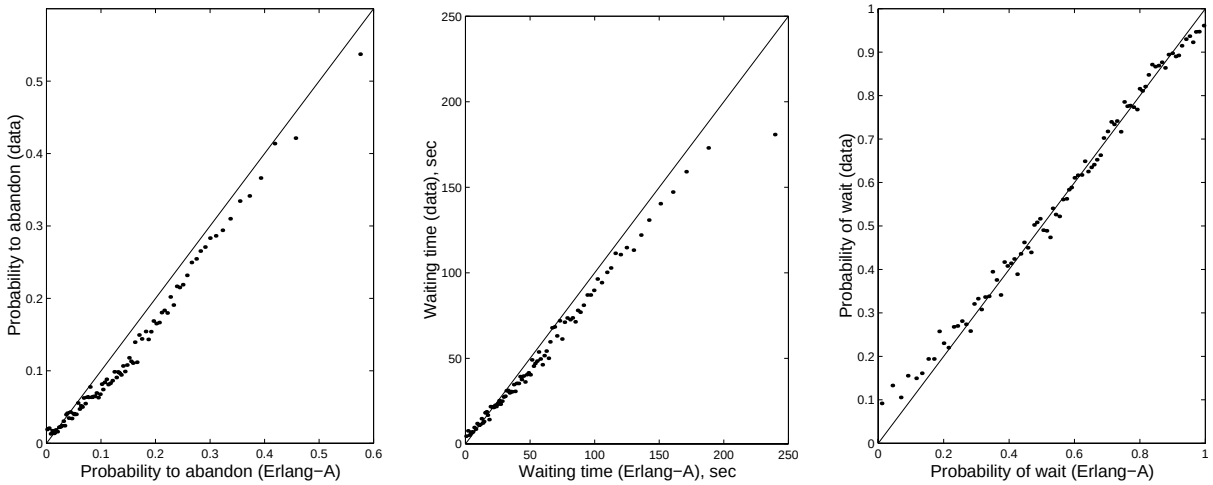
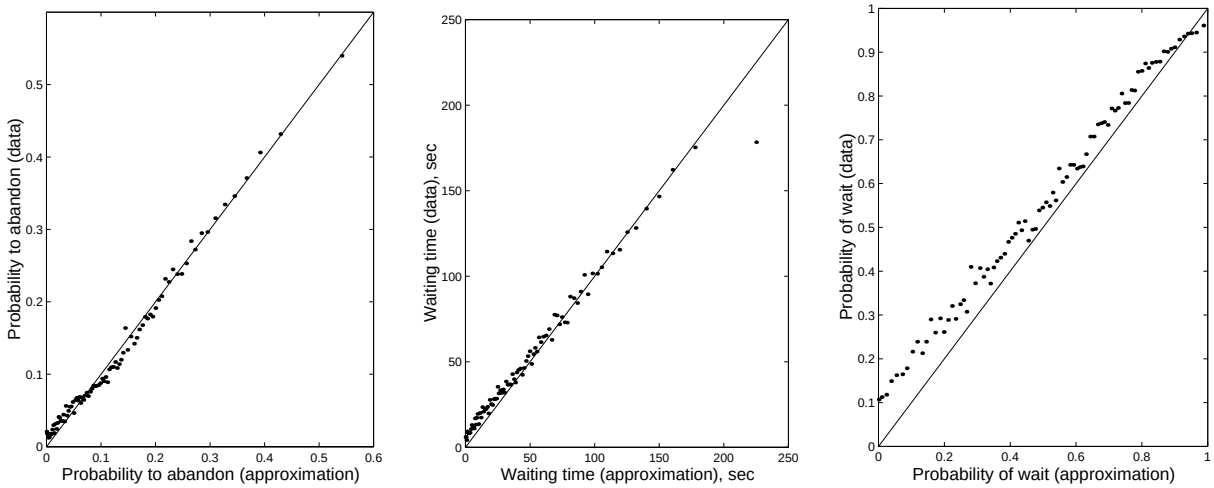


Figure 26: Erlang-A approximations vs. data averages



tion in Figure 25 and better understanding the relationship between Erlang-A formulae and their approximations.

8.2.3 Use and Limits of the Erlang-A Model

The empirical fit of the simple Erlang-A model and its approximation turns out to be very (perhaps surprisingly) accurate. Thus, for our call center—and those like it—use of the Erlang A for capacity-planning purposes could and should improve operational performance. Indeed, the model is beyond

typical current practice, and one aim in the current paper is to help change this state of affairs.

What call-center attributes are likely to make use of the Erlang-A model appropriate? Ideally, one would like to perform similar, empirical analysis of other centers to identify what drives the model's very good fit.

Until then, we may use queueing-theoretic results for insight on settings where the Elang-A model is likely to *not* be applicable. For example, it has been shown that in larger and more highly utilized systems, the form of the service-time distribution may affect waiting-time characteristics in unexpected ways (Schwartz and Mandelbaum 2002). Therefore, the Erlang A's good fit—despite violations of exponentiality—may be limited to smaller, less highly utilized centers. More broadly, models such as the Erlang A, which assume a homogenous customer population, cannot address multiple customer classes and priority schemes (Serfozo 1999), not to mention skills-based routing of calls (Garnet and Mandelbaum 1999; Mandelbaum and Stolyar 2002; Atar, Mandelbaum and Reiman 2002).

Thus, when faced with the classical modeling tradeoff between *simplicity* and *validity*, we have opted for Erlang-A simplicity. However, it is important to note that, as the complexity and efficiency of call-center operations continue to increase, more sophisticated and sensitive analysis will be required. Indeed, the type of detailed statistical analysis developed in this paper will become essential in models far more complex than the Erlang-A.

9 CONCLUSION

In this paper, we have analyzed a unique database of call-by-call data from a relatively small telephone call center. Even given its small size, the original data set included more than 1,200,000 calls, roughly 450,000 of which were from customers who wished to speak with an agent. The focus of our analysis was this set of 450,000 calls.

Our analysis was guided by queueing theory. Our call-by-call data allowed us to characterize queueing primitives, such as the arrival process (as inhomogenous Poisson with additional randomness in its arrival rate), the service-time distribution (as lognormal), and the distribution of customer impatience. We used these building blocks to develop additional tools that are useful in call-center management: theoretical and empirical patience indices, as well as prediction intervals for the offered load. Finally, we tested the robustness of several queueing-theoretic results. We found that a simple multi-server generalization of the classical Khintchine-Pollaczek formula produced biased waiting-time predictions. In contrast, queueing results concerning the (linear) relationship between average waiting times and abandonment rates, as well as predictions derived from the Erlang A (M/M/N+M) model, proved to be surprisingly robust.

The analysis of these data has also prompted us to develop new statistical methods and approaches. Specifically, our characterization of the service-time distribution gave rise to new, nonparametric methods for estimating regression models with lognormal errors. The large volume of highly censored abandonment data motivated us to develop nonparametric methods of estimating and graphing associated hazard-rate distributions.

Finally, we note that our analysis of the current data has generated a number of new questions. Statistical problems include the development of tools for survival analysis: how to analyze very large data sets with dependencies, and how to estimate means when given very high censoring rates. Queueing-theoretic problems include the understanding of the circumstances under which distributional assumptions, such as exponentiality, are important and when they are not.

References

Agresti, A. (1990), *Categorical Data Analysis*, John Wiley & Sons.

Aalen, O. O., and Gjessing, H. (2001), "Understanding the shape of the hazard rate: A process point of view", *Statistical Science*, **16**, 1–22.

Anscombe, F. (1948), "The transformation of poisson, binomial and negative-binomial data", *Biometrika*, **35**, 246–254.

Atar, R., Mandelbaum, A., and Reiman, M. (2002), "Scheduling a multi-class queue with many i.i.d. servers: asymptotic optimality in heavy traffic", *Working Paper*, Technion, Israel Institute of Technology.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Baccelli, F., and Hebuterne, G. (1981), "On queues with impatient customers", *International symposium on computer performance*, pp. 159–179.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society B*, **57**, 289–300.

Bolotin, V. (1994), Telephone circuit holding time distributions, *14th International Tele-traffic Conference (ITC-14)*, Elsevier.

Borst, S., Mandelbaum, A., and Reiman, M. (2000), "Dimensioning of large call centers", *Preprint*.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Breukelen, G. (1995), "Theorectial note: Parallel information processing models compatible with lognormally distributed response times", *Journal of Mathematical Psychology*, **39**, 396–399.

Brown, L. D., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2002), "Multifactor poisson and gamma-poisson models for call center arrival times", *Technical Report*.

Brown, L., and Hwang, J. (1993), "How to approximate a histogram by a normal density", *The American Statistician*, **47**, 251–255.

Brown, L., and Shen, H. (2002), "Analysis of service times for a bank call center data", *Technical Report*, University of Pennsylvania.

Brown, L., Zhang, R., and Zhao, L. (2001), "Root un-root methodology for nonparametric density estimation", *Technical Report*, University of Pennsylvania.

Brown, L., and Zhao, L. (2001), "A new test for the Poisson distribution", *To appear in Sankhya*.

Buffa, E., Cosgrove, M., and Luce, B. (1976), "An integrated work shift scheduling system", *Decision Sciences*, **7**, 620–630.

Buzacott, J., and Shanthikumar, J. (1993), *Stochastic models of manufacturing systems*, Prentice Hall.

Call Center Statistics (2002), Available from Call Center News Service Web Site, `http://www.callcenternews.com/resources/statistics.shtml`.

Cappe, O., Moulines, E., Pesquet, J. C., Petropulu, A. P., and Yang, X. "Long-range dependence and heavy-tail modeling for teletraffic data", *IEEE Signal Processing Magazine*, **19**, 14–27.

Dette, H., Munk, A., and Wagner, T. (1998), "Estimating the variance in nonparametric regression–what is a reasonable choice?", *Journal of Royal Statistical Society, Series B*, **60**, 751–764.

Erlang, A. (1911), "The theory of probability and telephone conversations", *Nyt Tidsskrift Mat. B*, **20**, 33–39.

— (1917), "Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges", *Electroteknikeren (Danish)*, **13**, 5–13. English translation 1918 P.O. Elec. Eng. J. 10, 189–197.

Fendick, K., Saksena, V., and Whitt, W. (1989), "Dependence in packet queues", *IEEE Transactions on Communications*, **37**, 1173–1183.

Gans, N., Koole, G., and Mandelbaum, A. (2002), "Telephone calls centers: a tutorial and literature review", *Technical Report*.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Garnett, O., and Mandelbaum, A. (1999), *An Introduction to Skills-Based Routing and its Operational Complexities*, Teaching note sponsored by the Fraunhofer IAO Institute, Stuttgart, Germany.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Garnett, O., Mandelbaum, A., and Reiman, M. (2002), "Designing a call-center with impatient customers", *To appear in MSOM*.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Hall, P., Kay, J., and Titterington, D. (1990), "Asymptotically optimal difference-based estimation of variance in nonparametric regression", *Biometrika*, **77**, 521–528.

Hall, R. (1991), *Queueing methods for services and manufacturing*, Prentice Hall.

Harris, C., Hoffman, K., and Saunders, P. (1987), "Modeling the IRS telephone taxpayer information system", *Operations Research*, **35**, 504–523.

Iglehart, D., and Whitt, W. (1970), "Multiple channel queues in heavy traffic, i and ii", *Advances in Applied Probability*, **2**, 150–177, 355–364.

iProfiler Software (2002), Available at 4CallCenters Web Site, `http://www.4CallCenters.com`.

Jongbloed, G., and Koole, G. (2001), "Managing uncertainty in call centers using poisson mixtures", *Applied Stochastic Models in Business and Industry*, **17**, 307–318.

Kingman, J. F. C. (1962), "On queues in heavy traffic", *Journal of the Royal Statistical Society B*, **24**, 383–392.

Kooperberg, C., Stone, C., and Truong, Y. (1995), "Hazard regression", *Journal of the American Statistical Association*, **90**, 78–94.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, Second Edition, Chapman and Hall, New York and London.

Levins, M. (2002), "On The New Local Variance Estimator", *Working PhD Thesis*, University of Pennsylvania.

Loader, C. (1999), *Local Regression and Likelihood*, Springer-Verlag, New York.

Mandelbaum, A. (2001), "Call centers. research bibliography with abstracts", *Technical report*, Technion, Israel Institute of Technology.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Mandelbaum, A., Sakov, A., and Zeltyn, S. (2000), "Empirical analysis of a call center", *Technical report*, Technion, Israel Institute of Technology.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Mandelbaum, A., and Schwartz, R. (2002), "Simulation Experiments with M/G/100 Queues in the Halfin-Whitt (Q.E.D) Regime", *Technical Report*, Technion.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Mandelbaum, A., and Stolyar, A.L. (2002), "Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule", *Working Paper*, Technion, Israel Institute of Technology.
Downloadable from http://iew3.technion.ac.il/serveng/References/references.html.

Müller, H., and Stadtmüller, U. (1987), "Estimation of heteroscedasticity in regression analysis", *The Annals of Statistics*, **15**, 610–625.

Palm, C. (1943), "Intensitatsschwankungen im fernsprechverkehr", *Ericsson Technics*, **44**, 1–189.

Palm, C. (1953), "Methods of judging the annoyance caused by congestion", *Tele*, **4**, 189–208.

Serfozo, R. (1999), *Introduction to Stochastic Networks*, Springer-Verlag, New York.

Shen, H. (2002), "Estimation, Confidence Intervals and Nonparametric Regression for Problems Involving Lognormal Distribution", *Working Phd Thesis*, University of Pennsylvania.

Sze, D. (1984), "A queueing model for telephone operator staffing", *Operations Research*, **32**, 229–249.

Uchitelle, L. (2002), "Answering '800' calls, extra income but no security", *The New York Times*, **March 27**, Section A, Page 1, Column 5.

Ulrich, R., and Miller, J. (1993), "Information processing models generating lognormally distributed reaction times", *Journal of Mathematical Psychology*, **37**, 513–525.

Whitt, W. (1993), "Approximations for the GI/G/m queue", *Production and Operations Management*, **2**, 114–161.

— (2002), *Stochastic-Process Limits*, Springer-Verlag.

Willinger, W., Taqqu, M. S., Leland, W. E., and Wilson, D. V. (1995), "Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements", *Statistical Science*, **10**, 67–85.

Wolff, R. (1989), *Stochastic Modeling and the Theory of Queues*, Prentice Hall.

Zohar, E., Mandelbaum, A., and Shimkin, N. (2002), "Adaptive Behavior of Impatient Customers in Tele-queues: Theory and Empirical Support", *Management Science*, **48(4)**, 556–583.